Niveau(x) d'enseignement: 6 Numéro EDUQ : 0870466
Centre d'intérêt: 090

Colloque Recherches sur l'apprentissage : analyse des correspondances et méthodes statistiques apparentées (1986 : Université du Québec à Montréal).

Recherches sur l'apprentissage : analyse des correspondances et méthodes statistiques apparentées / 26, 27 et 28 mai 1986 ; colloque organisé par le CIRADE en collab. avec le SCAD de l'Université du Québec à Montréal. -- [Montréal] : UQAM, CIRADE, [1986?].
135, 5 p. : ill. ; 28 cm.

"Avec l'appui financier de la Communauté scientifique Réseau et du PAFACC". Comprend des bibliogr.

Vedette(s)-secondaire(s): Université du Québec à Montréal. Centre interdisciplinaire de recherches sur l'apprentissage et le développement en éducation * Université du Québec à Montréal. Service de consultation en analyse des données

Descripteur(s): Recherche en éducation * Apprentissage * Méthode de recherche * Statistique * Analyse de données * Universitaire (Ordre d'enseignement)

Ce colloque visait à mieux comprendre les diverses méthodes exploratoires d'analyse de données, les principes sous-jacents qui les guident et l'utilisation qui en est faite. Les textes des quatre communications faites au colloque, présentés dans le document, incluent une partie théorique situant les fondements mathématiques de ces méthodes, et une partie plus pratique appuyée par des exemples. Le colloque a permis d'autre part de réfléchir, sur la base d'exemples précis, aux apports respectifs et aux limites de ces méthodes. Les synthèses des deux ateliers et de la table ronde reprennent quelques éléments de cette réflexion. Ce colloque fut l'occasion de favoriser un échange sur des méthodes de recherche utilisées par différents chercheurs, et d'amorcer une réflexion sur les orientations actuelles des recherches sur l'apprentissage et sur les problèmes de méthodologie qu'elles posent.

Type de document: f

EDUQ 8870466

CENTRE INTERDISCIPLINAIRE DE RECHERCHES SUR
L'APPRENTISSAGE ET LE DEVELOPPEMENT EN EDUCATION

cirade

Université du Québec à Montréal

EDUQ 8870466

"Recherches sur l'Apprentissage: Analyse des correspondances et méthodes statistiques apparentées"

26, 27 et 28 mai 1986

CIRADE - UQAM

Colloque organisé par le CIRADE (centre interdisciplinaire de recherche sur l'apprentissage et le développement en éducation), en collaboration avec le SCAD (service de consultation en Analyse des données) de l'Université du Québec à Montréal; avec l'appui financier de la Communauté scientifique Réseau et du PAFACC.

TABLE DES MATIERES

Page

Introduction 1

1) Résumé de la présentation de Régis Gras, situant la problématique sous-jacente du colloque 3

2) Conférence de Mme Brigitte Escofier Université de Rennes : La méthode d'analyse des correspondances : grandes lignes théoriques de la méthode, utilisation et dépouillement des résultats 7

3) Conférence de M. Régis Gras, Université de Rennes : Deux méthodes d'analyse de données didactiques : Classification implicative et classification hiérarchique. 23

4) Communication de M. Richard Bertrand, Université Laval : Méthode de régression en analyse exploratoire 45

5) Communication de Bernard Clément, Ecole Polytechnique, Montréal : Un programme SAS pour faire de l'analyse des correspondances 66

6) Compte rendu de l'Atelier 1 : Atelier d'analyse de données par la méthode d'analyse des correspondances 84

7) Compte rendu de l'Atelier 2 :
Atelier d'analyse de données didactiques par la méthode de classification hiérarchique 100

8) Compte rendu de la table ronde :
Questions reliées aux recherches actuelles sur l'apprentissage et aux méthodologies de recherche appropriées 115

9) Annexes :
1. Programme du colloque

2. Liste des documents disponibles au CIRADE (suite au colloque).

INTRODUCTION

Pour tous les projets du centre, les éléments méthodologiques revêtent une importance capitale. Cherchant à décrire et comprendre, dans toute leur complexité, les processus par lesquels les conceptions et procédures de l'apprenant peuvent changer et évoluer en situation, ils doivent souvent faire appel à des démarches "non-classiques" d'analyse de données.

Ce colloque correspond donc à une préoccupation fondamentale des chercheurs-es du centre et de plusieurs chercheurs-es actuels-les en éducation. Les orientations récentes des recherches sur l'apprentissage tendent en effet beaucoup plus à clarifier, pour mieux comprendre, les phénomènes en présence qu'à prouver ou confirmer des hypothèses établies à priori.

Dans cette conjoncture, certaines méthodes exploratoires d'analyse de données peuvent s'avérer très intéressantes pour plusieurs chercheurs-es en éducation. La méthode d'analyse factorielle des correspondances, les méthodes classificatoires et implicatives constituent, en ce sens, et avec d'autres, des supports importants dans le traitement exploratoire de données de recherche impliquant plusieurs variables. Cependant, ces méthodes, à l'origine françaises, sont assez mal connues autant par les chercheurs-es qui les utilisent que par les statisticiens-nes appelés-es à les traiter.

Le présent colloque visait à mieux comprendre ces différentes méthodes, les principes sous-jacents qui les guident (modèles mathématiques...) et l'utilisation qui en est faite. Il rejoignait ici autant la préoccupation des chercheurs-es en éducation que des statisticiens-es. Les textes des conférences présentées dans les actes du colloque incluent ainsi une partie théorique situant les fondements mathématiques de ces méthodes, et une partie plus pratique appuyée par des exemples.

Le colloque permettait d'autre part de réfléchir, sur la base d'exemples précis, aux apports respectifs et aux limites de ces méthodes. Les textes des ateliers et de la table ronde reprennent quelques éléments de cette réflexion.

Enfin, il était l'occasion de favoriser un échange sur des méthodes de recherche utilisées par différent(e)s chercheur(e)s, et d'amorcer ainsi une réflexion sur les orientations actuelles des recherches sur l'apprentissage et sur les problèmes méthodologiques qu'elles posent.

Nadine Bednarz
Responsable du colloque

Juin 1986

Nous tenons à remercier tout particulièrement Jacqueline Clément et Guylaine St-Pierre pour l'aide précieuse qu'elles ont apportée dans la dactylographie des ces actes et Sophie Renée de Cotret pour son aide dans la lecture et finalisation des documents.

DEUX METHODES D'ANALYSE DE DONNEES DIDACTIQUES

INTRODUCTION

par Régis Gras, Institut de Recherche Mathématique de l'Université de Rennes, France.

Problématique

La statistique classique centre son étude sur les tests et sur l'estimation de paramètres. Elle **nécessite** de ce fait, en général, **la présomption et l'émission d'hypothèses,** en vue de les valider ou de les refuter, que ces hypothèses s'expriment en termes de valeurs possibles ou bien de lois a priori ou bien d'homogénéité de population.

Or nous nous trouvons aujourd'hui dans une conjoncture différente de celle de 1900 ou même de 1940.

1) d'une part, les données fournies par les observations sont de plus en plus nombreuses et variées (dans leur nature). Leur enregistrement s'entoure d'une certaine scientificité;

2) d'autre part, les ordinateurs et même les calculatrices programmables ouvrent une ère où ces données peuvent être prises en compte et traitées, alors que 50 ans plus tôt, elles auraient été mises au rebut car inexploitables;

3) enfin, le champ de la statistique "moderne" s'élargit de plus en plus en sciences humaines et, particulièrement pour nous, en didactique. Dans ces domaines, les hypothèses sont difficilement formulables et le coût de l'homogénéisation des populations est disproportionné par rapport à la fiabilité de celle-ci.

La convergence de ces faits nous conduit à une **démarche nouvelle.** Les **données** didactiques, par exemple, peuvent être **étudiées sans a priori** et, surtout, **globalement,** contrairement aux méthodes classiques qui nécessitent des examens séquentiels des caractéristiques. Une approche multidimensionnelle devient possible et mène à la **mise en évidence de structures** et de **facteurs latents,** inapparents ou indiscernables dans

l'examen brut des données, quitte à ce que cette connaissance conduise à la formulation d'hypothèses dont la statistique classique pourrait se charger.

Un autre caractère de la démarche nouvelle est la **recherche systématique de la visualisation des phénomènes** : nuages de l'analyse factorielle, noyaux de partition dans les nuées dynamiques, arbres de classification, etc... Elle nécessite le fonctionnement de théories mathématiques de géométrie, d'algèbre linéaire, de mécanique classique, de combinatoire, etc...

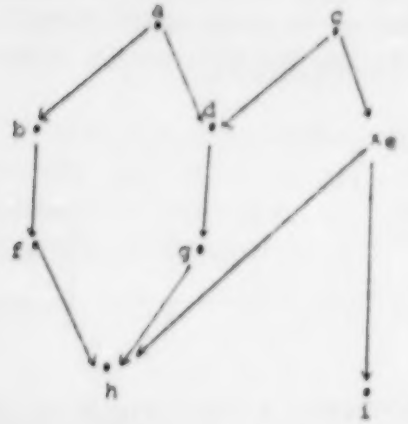
Méthodologie propre aux méthodes de classification hiérarchique et de classification implicative

Visant à mettre en évidence une certaine similarité entre comportements ou attributs, elles s'appuient sur la même idée : cette similarité est d'autant plus grande qu'elle paraît "invraisemblable" sous une hypothèse d'absence de lien entre les comportements observés ou les attributs possédés. La construction hiérarchique selon un arbre permet l'analyse des données initiales en termes structuraux. Ces idées bien vagues vont se préciser à travers chacune des méthodes développées.

Classification implicative

Un problème commun à la psychologie et à la didactique, mais aussi à de nombreux autres domaines (archéologie par exemple) a présidé à la naissance de cette méthode : est-il possible d'ordonner en un préordre partiel sinon total des comportements binaires (ou binarisés) selon la relation : "le comportement a entraîne le comportement b"?

Un nouvel indice, auquel on peut donner plusieurs formes, construit sur des bases probabilistes et statistiques, permet de répondre à cette question, modulo un seuil maîtrisable à volonté. Il suffit d'examiner l'effectif des individus "contredisant" l'implication : $a \Rightarrow b$ et de le comparer à celui qui aurait été obtenu dans l'hypothèse d'absence de subordination de b à a. Pour un seuil donné, on en dégage un tableau d'implication entre les comportements observés puis un graphe transitif du type suivant :

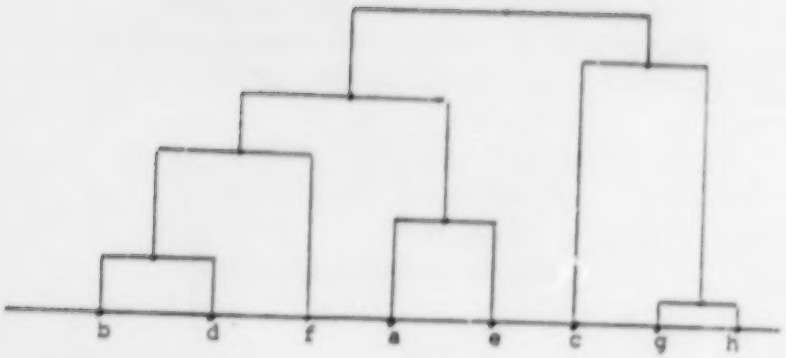


L'examen de la structure du graphe permet d'émettre, dans le cas didactique, des hypothèses de dépendance et d'identifier les chaînes en des termes de complexité et de nature des tâches. Cette méthode, toute récente, vient d'être utilisée dans un test de maîtrise du concept de symétrie centrale par des élèves de 4ème, les comportements retenus étant la réussite ou l'échec aux items du test.

Classification hiérarchique (I.C. Lerman)

Cette méthode accepte le traitement de données très diverses : non seulement des variables binaires, attributs ou comportements, mais également des variables ordinales ou préordinales. Ainsi, un questionnaire d'attitudes à l'égard des mathématiques, présentant des questions à modalités ordonnées ("beaucoup", "assez", "peu",...) a pu être analysé par ce procédé.

Par exemple, la similarité entre deux attributs sera jugée d'autant plus forte que l'effectif des individus les possédant simultanément est grand par rapport à celui observable dans l'hypothèse d'une absence de lien entre ces attributs. Un indice prolongeant celui-ci dans le cas d'attributs, d'ordres, de préordres ou de classes d'ordres, se définit, selon la même démarche, sur des bases probabilistes et combinatoires. Chaque indice conduit à un arbre hiérarchique du type suivant :



Chaque noeud correspond à une variable ou une classe de variables. A chaque niveau horizontal correspond une partition dont une statistique souligne la pertinence. L'analyse consiste à donner un sens à cette partition.

De nombreux tests et questionnaires traités par la méthode ont fourni des résultats remarquables dans la connaissance de phénomènes cognitifs et affectifs.

Quelques problèmes:

• de type technique : le traitement des données didactiques nécessite un ordinateur dès que la taille de l'ensemble des variables devient importante (15 à 20 par exemple) et que celle des individus dépasse deux classes scolaires. Une calculatrice programmable, type TI 58 ou 59, permet de calculer les indices nécessaires. Cependant, des programmes peu coûteux existent et permettent dans chaque université de faire des études sur des échantillons volumineux. Une ombre demeure : il n'existe pas encore de construction automatique d'arbre implicatif.

• de type didactique : les hypothèses formulables à partir d'une analyse demandent des confirmations à travers d'autres expériences où l'on ferait varier les outils d'analyse (tests, méthodes,...) et les paramètres définissant les situations didactiques. Une analyse n'a ni la possibilité ni la prétention de tout dire et la recherche d'invariants nécessite de multiplier les approches.

L'ANALYSE DES CORRESPONDANCES: grandes lignes théoriques de la méthode, utilisation et dépouillement des résultats

par Brigitte Escofier, Institut de Recherche en Informatique et Systèmes aléatoires de l'Université de Rennes, France.

1- INTRODUCTION

Commençons par un très bref historique. L'analyse des correspondances est une méthode d'analyse des données "inventée" par J.P. Benzécri, et programmée pour la première fois à Rennes en 1964. Indépendamment, d'autres auteurs ont introduit des formules analogues, mais Benzécri avait senti la puissance de cette méthode et après avoir fait écrire le premier programme, l'avait testé systématiquement sur des données les plus diverses provenant de tous les domaines.

Maintenant, presque 25 ans après, comme le bon vin, elle a bien vieilli, s'est enrichie en puissance et en qualité. Elle est devenue l'élément essentiel des chaires de programmes d'analyse des données; que ce soit dans sa forme première, analyse des corresponsances simples ou dans sa généralisation, l'analyse des correspondances multiples. Cette dernière, permet notamment, de dépouiller les résultats d'enquêtes en synthétisant l'ensemble des réponses de milliers d'individus à un grand nombre de questions.

Je vais expliquer le principe, les formules, les principaux résultats de l'Analyse des Correspondances et son utilisation à partir d'un petit exemple linguistique. Ceci est encore un retour aux sources puisque l'idée initiale de Benzécri était le traitement de ce type de données.

2- LE TABLEAU DE DONNEES

L'exemple retenu ici pour des fins explicatives est de petite taille, ce qui permet de le dépouiller facilement. Il ne prétend pas du tout résoudre des problèmes linguistiques ou politiques, ce n'est qu'une illustration de la méthode d'analyse des correspondances.

J'ai pris, ce qu'on appelle, les déclarations d'intention de candidats aux élections législatives de 1978. Ce sont des textes d'une à deux pages, signés par les candidats et distribués à tous les électeurs quelques jours avant la date des élections. Dans le canton de Rennes-Nord, 12 candidats se présentaient, représentant des partis politiques assez variés. Pour procéder à l'analyse de ces textes, on a choisi 27 mots jugés caractéristiques et surtout assez fréquents dans l'ensemble des 12 textes. Puis, on a procédé à un comptage et on a construit un tableau à 27 lignes et 12 colonnes contenant au croisement de la ligne i (correspondant au mot i) et de la colonne j (correspondant au candidat j) le nombre de fois que le mot était utilisé dans le texte.

L'analyse des correspondances (simple) commence toujours par la construction d'un tel tableau, qu'on appelle tableau de fréquences, de contingence. Plus généralement c'est un tableau qui contient des nombres positifs ou nuls représentant des comptages.

Par la suite, très souvent, au lieu de considérer le tableau de comptage, on considérera un tableau déduit du précédent en divisant chacun des termes Kij par leur somme. Ce nouveau tableau définit une loi de probabilité sur le produit I x J.

On définit les marges des tableaux. Ce sont les sommes des lignes ou des colonnes. La marge sur I est notée k_i ou p_i suivant que l'on travaille sur le tableau de fréquences ou de probabilité. Dans l'ensemble, K_i représente le nombre total d'utilisations du mot i, et p_i sa probabilité marginale. On note de même k_j et p_j la marge sur J; le premier est le nombre total de mots utilisés dans le texte par le candidat j, tandis que p_j est la probabilité marginale.

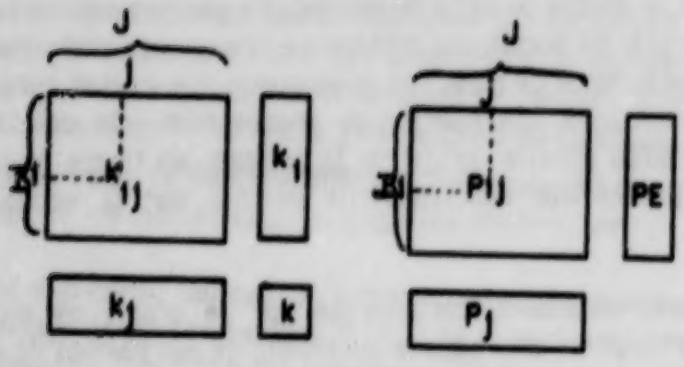


Tableau de comptage — Probabilité

Effectif total: $k = \sum_{i,j} k_{ij}$

$p_{ij} = k_{ij}/k.$ $\quad \sum_{i,j} p_{ij} = 1$

Marges sur I : $k_i = \sum_j k_{ij}$ $\quad p_i = \sum_j p_{ij}$

Marges sur J : $k_j = \sum_i K_{ij}$ $\quad p_j = \sum_i p_{ij}$

Figure 1 : Les tableaux et leurs marges

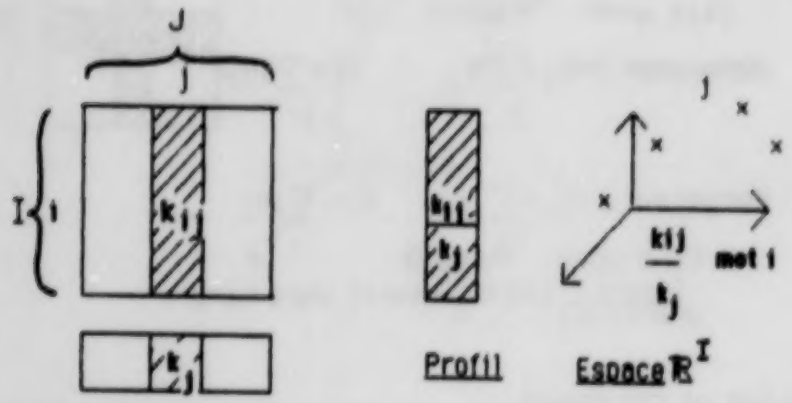
3- LE NUAGE DE COLONNES

L'étude de ce tableau a notamment pour but de mettre en évidence les différences de vocabulaire entre les candidats. La colonne associée à un candidat contient la fréquence d'utilisation des 27 mots sélectionnés. La ressemblance entre les colonnes du tableau va nous permettre de définir une notion de proximité entre les candidats, basée sur leur vocabulaire.

Si on comparait directement, ligne par ligne, les nombres contenus dans 2 colonnes, le nombre total de mots utilisés par chacun des candidats jouerait un rôle essentiel. Si deux candidats utilisaient le même vocabulaire de la même façon, mais que l'un était deux fois plus bavard que l'autre, la colonne du

premier serait le double de celle du second. En gardant comme caractéristique les nombres bruts, on aurait une différence importante entre eux; ce qui n'est pas souhaitable puisque ce n'est pas la longueur des textes qui nous intéresse mais les pourcentages d'utilisation de chaque mot. On considère donc ces pourcentages, qu'on appelle profil de la colonne; en terme de probabilités il s'agit de la probabilité conditionnelle définie sur le vocabulaire par le candidat.

Chacun des candidats, caractérisé par le profil de sa colonne est représenté par un point d'un espace euclidien; ce qui permet de visualiser les proximités entre eux. Cet espace euclidien a une dimension égale au nombre de lignes du tableau, ici 27. Le candidat j est représenté par un point dont la coordonnée sur l'axe i (associée au mot i) est le pourcentage k_{ij}/k_j (voir figure 2). Chacune des 12 colonnes est ainsi représentée dans $\mathbb{R}^{27}$, on a donc un nuage de 12 points. Plus les profils des colonnes se ressemblent, plus les points sont proches, à la limite deux colonnes proportionnelles ont deux points confondus.



Le candidat j est caractérisé par le profil de sa colonne et représenté par un point de l'espace $\mathbb{R}^I$

Figure 2 : Le nuage des colonnes

4- LE NUAGE DES LIGNES

En construisant le nuage des 12 candidats, on a considéré l'ensemble des colonnes du tableau. Une construction absolument analogue peut être faite en considérant les lignes de ce tableau. En effet, dans un tableau de comptage

l'ensemble des lignes et celui des colonnes ont un rôle symétrique. Chaque mot est donc caractérisé par une ligne du tableau, ou plus exactement par le profil de cette ligne. Il est représenté par un point de l'espace $\mathbb{R}^{12}$, puisqu'il y a 12 colonnes. La coordonnée sur l'axe j correspond au candidat j en k_{ij}/k_i. Deux mots sont proches si ce sont les mêmes candidats qui les utilisent beaucoup tous les deux, et les mêmes candidats qui les utilisent peu ou pas du tout.

Partant d'un même tableau, on a donc construit un nuage de 12 points dans $\mathbb{R}^{27}$ et un nuage de 27 points dans $\mathbb{R}^{12}$. Ces nuages sont les représentations d'un même tableau, l'un à travers ses lignes, l'autre à travers ses colonnes. De très fortes relations existent donc entre eux, on les appelle relations de dualité.

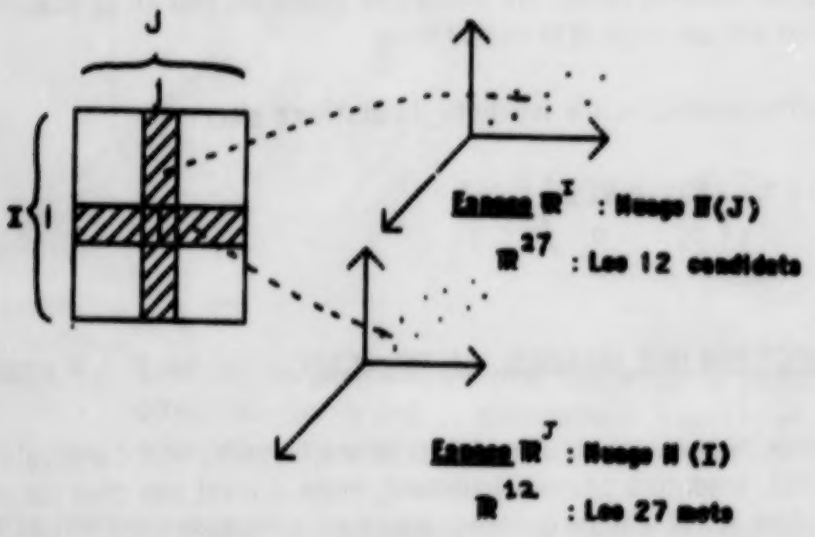


Figure 3 : Les 2 représentations duales d'un même tableau

5- LA DISTANCE DU χ^2 (Chi carré)

En analyse des correspondances la distance entre 2 points est une distance particulière, la distance du χ^2. La distance classique est la somme des carrés des différences des coordonnées; dans la distance du χ^2 un terme pondère cette somme. La distance entre les profils de deux colonnes j et j' s'écrit:

$$D^2(j,j') = \sum_i \left(\frac{p_{ij}}{p_j} - \frac{p_{ij'}}{p_{j'}} \right)^2 \frac{1}{p_i}$$

Le terme 1/Pi équilibre l'influence de chacun des mots sur la distance entre 2 candidats et permet d'avoir une propriété particulière, "l'équivalence distributionnelle". Cette expression traduit la propriété suivante: si dans un tableau, deux lignes exactement proportionnelles sont cumulées en une seule, la distance entre les profils des colonnes est inchangée. (Si deux mots s'avèrent équivalents, leur regroupement ne change pas les distances entre candidats). Le cas de proportionnalité exacte ne se rencontre pas, mais dans le cas de grande ressemblance, les distances changent peu et la stabilité de la construction est une propriété essentielle.

Symétriquement, entre les mots, la distance s'écrit :

$$D^2(i,i') = \sum_j \left(\frac{p_{ij}}{p_i} - \frac{p_{i'j}}{p_{i'}} \right)^2 \frac{1}{p_j}$$

6- PROJECTION DES NUAGES: LE PRINCIPE

Les deux nuages construits, les proximités entre les candidats (et les mots) ont été traduites géométriquement, mais il n'est pas plus facile de les saisir que dans le tableau de données, puisque ces nuages sont situés dans des espaces de grande dimension dans lesquels nous ne pouvons les visualiser. Ainsi, nous allons projeter ces nuages sur des plans pour représenter graphiquement la proximité entre profils. On traduit ainsi visuellement l'essentiel de la structure définie par le tableau de chiffres.

Dans un graphique plan, on ne peut généralement pas avoir une représentation exacte du nuage de points, à moins que celui-ci ne soit complètement plan. On obtient donc une représentation approchée de ce nuage.

Pour déformer le moins possible le nuage, il faut bien choisir les plans sur lesquels il sera projeté.

Tout d'abord l'origine des axes dans l'espace est placée au centre de gravité G du nuage. Pour définir le centre de gravité, il faut préciser le poids

de chaque point. En analyse des correspondances le poids du point j est fixé à p_j. Le centre de gravité du nuage est alors le point de coordonnée pi; c'est la marge sur I.

Ensuite on cherche la direction d'allongement la plus importante du nuage. En effet (cf. figure 4) la projection dans une direction de grand allongement du nuage le déformera moins que la projection dans une direction de faible dispersion.

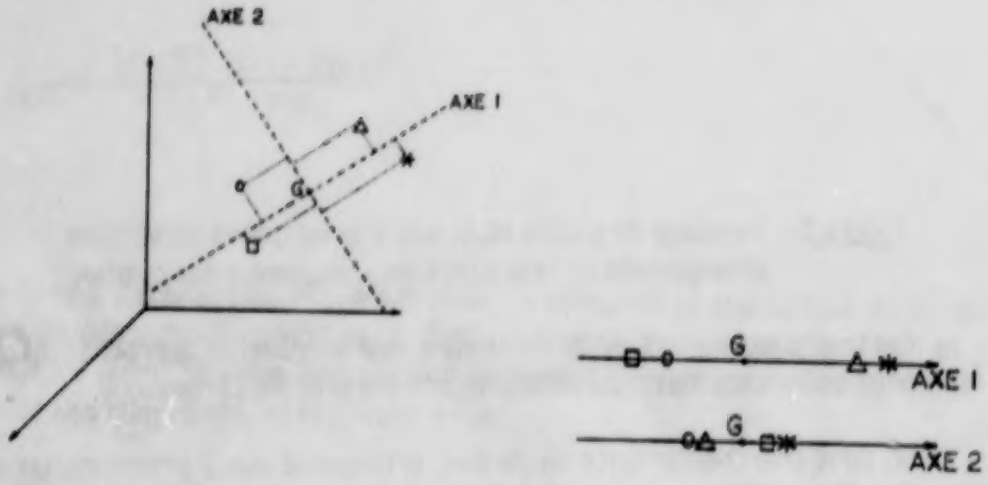


Figure 4 : Exemple de projection d'un nuage de 4 points dans une direction de grand allongement (axe I) et de faible allongement (axe 2).

Cette première direction d'allongement est l'axe I. La projection de tous les points du nuage est appelée le premier facteur.

Ensuite on cherche un deuxième axe, passant par le centre de gravité, orthogonal au premier qui correspond encore à la direction d'allongement maximal. La projection sur ces 2 axes orthogonaux nous donne une projection sur un plan (cf figure 5)

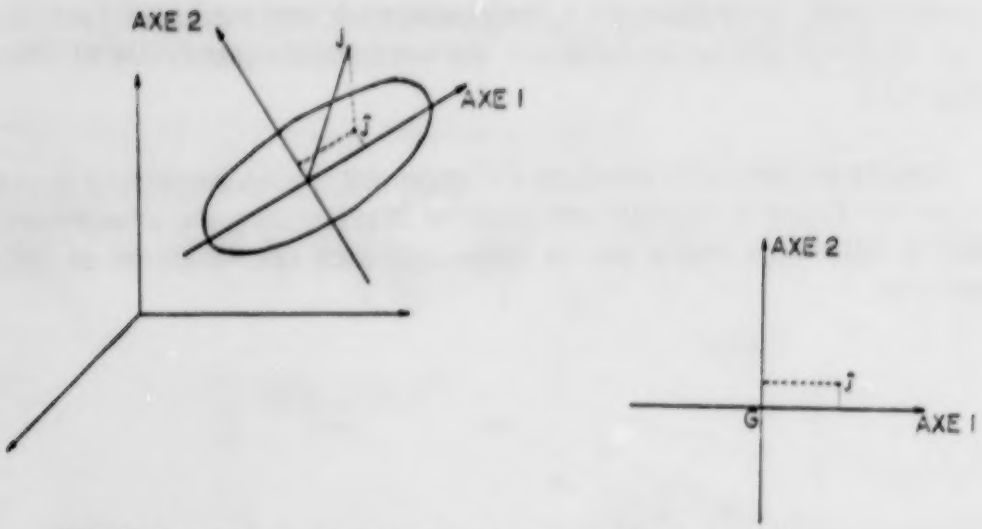


Figure 5 : Le nuage de points N(J), ses 2 principales directions allongement et la projection $\hat{j}$ du point j sur le plan.

Cette projection donne la meilleure représentation approchée plane des distances entre candidats, au sens d'un critère précisé ci-dessous.

On peut chercher ensuite un 3e axe, orthogonal aux 2 premiers, un 4e axe etc... On obtient des représentations planes avec les axes 2-3, 3-4, etc...

7- L'INERTIE MAXIMALE. CRITERE DE PROJECTION

Le critère utilisé pour définir la direction d'allongement maximal est un critère d'inertie. L'inertie d'un point j par rapport au centre de gravité G du nuage est le produit de son poids P_j par le carré de sa distance à G. L'inertie du nuage est la somme des inerties de tous les points.

Le premier axe est, par définition, l'axe sur lequel le nuage projeté a une inertie maximale. Le deuxième axe étant orthogonal au premier, l'inertie du nuage projeté sur le plan 1-2 est la somme des inerties des axes 1 et 2, etc.

8- INERTIE DES DEUX NUAGES

Le nuage des lignes se traite exactement de la même façon que le nuage des colonnes. Son centre de gravité est confondu avec la marge P_j.

La "dualité" entre ces deux nuages, qui sont issus du même tableau se traduit notamment en terme d'inertie. L'inertie du nuage des colonnes s'écrit.

$$\text{Inertie de } N(J) = \sum_J \text{poids} \times [\text{ distance de } j \text{ à } G\ ]^2$$

$$= \sum_J p_j \sum_I \left(\frac{p_{ij}}{p_j} - p_i\right)^2 \frac{1}{p_i}$$

$$= \sum_{I,J} \frac{(p_{ij} - p_i p_j)^2}{p_i p_j}$$

Cette dernière expression est, à $1/k$ près, la valeur de la statistique du χ^2 qui mesure l'indépendance entre I et J. Cette expression est symétrique en i et j. L'inertie du nuage N(I) que l'on obtient en permutant i et j est donc la même : L'inertie des nuages N(I) et N(J) vaut χ^2/k.

9- ANALYSE DE CORRESPONDANCE ET INDEPENDANCE

Rappelons qu'en terme de probabilité, il y a indépendance si pour tout i et tout j, on a l'égalité $p_{ij} = p_i p_j$. Dans ce cas le profil p_{ij}/p_j de la colonne j est égal à la marge p_i : tous les points du nuage des colonnes sont confondus avec leur centre de gravité.

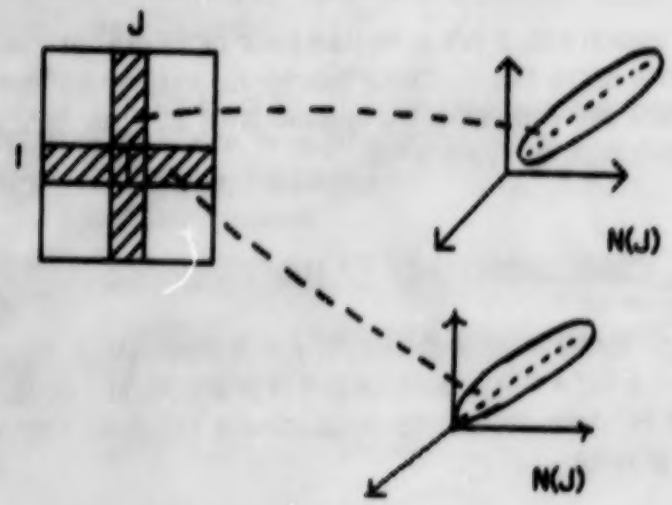
De même les profils des lignes p_{ij}/p_i sont égaux à la marge p_j. Le nuage des lignes est concentré sur son centre de gravité.

L'analyse des correspondances en analysant la dispersion de ces deux nuages analyse l'écart du tableau p_{ij} à l'hypothèse d'indépendance. L'inertie totale donne une mesure de cet écart. Les projections sur les axes indiquent les lignes et les colonnes les plus éloignées de cette hypothèse (c'est-à-dire de la marge).

10- PROJECTION DES DEUX NUAGES : DUALITE

Nous avons vu au paragraphe 8 que l'inertie des 2 nuages, lignes et colonnes était la même. Mais la dualité est beaucoup plus forte : l'inertie des deux nuages sur chacun des axes de projection est la même.

De plus, les projections des lignes et des colonnes sont liées par des formules, appelées formules de transition. Si on note F_l (i) la projection de la ligne i sur l'axe l d'ordre l du nuage N(I), G_l(j) la projection de la colonne j sur l'axe l d'ordre l du nuage N(J), λ_l la valeur de l'inertie de la projection de ces deux nuages, on a les relations indiquées dans la figure 6.



F_l(i) = Projection i sur axe l

G_l(j) = Projection j sur axe l

$$F_l(i) = \frac{1}{\sqrt{\lambda_l}} \sum_j \frac{p_{ij}}{p_i} G_l(j)$$

$$G_l(j) = \frac{1}{\sqrt{\lambda_l}} \sum_i \frac{p_{ij}}{p_j} F_l(i)$$

Figure 6 : Formules de "transition"

Il est pratique de superposer les projections des 2 nuages correspondants aux mêmes axes. Sur ces graphiques on visualise simultanément les proximités entre les différents candidats et les proximités entre les mots. Les relations entre les projections des candidats et des mots ne s'expriment pas en termes de proximité mais en termes de centre de gravité : un candidat j est situé (à $1/\sqrt{\lambda_1}$ près) au centre de gravité des mots qu'il emploie, le poids de chaque mot i étant égal à k_{ij}. Attiré par les mots lourds, (qu'il emploie beaucoup), ce candidat sera situé dans leur zone, et loin des mots légers (qu'il emploie peu). Réciproquement les mots seront attirés par les candidats qui les emploient beaucoup et repoussés par les candidats qui les emploient peu.

II- EXEMPLE : LES DONNEES

Les listings commentés proviennent du logiciel ADDAD. Les 27 mots tirés des textes des candidats sont:

Changement, progrès, droit, liberté, indépendance, démocratie, dictature, travailleur, femme, salaire, prix, emploi, droite, gauche, famille, contre, crise, socialisme, communisme, autogestion, lutte, S.M.I.G., nucléaire, régime, capitalisme, programme commun.

Ils apparaissent en ligne et sont notés avec 4 caractères. Les candidats apparaissent en colonne, avec 4 caractères aussi. Les partis de ces candidats étaient respectivement.

LAFO	Démocratie prolétarienne
LECL	Démocratie chrétienne
HERV	Parti socialiste
PORT	Ligue communiste
GUIL	Lutte ouvrière
VANN	P.F.N. (droite)
CRES	U.D.R. (Majorité)

MART Autogestion P.S.U.

ROLL Parti communiste

DECA P.R. (majorité)

TECH Marxiste léniniste

GORV Démocrate Breton

NOMJ(J)	LAFO	LECL	HERV	PORT	QUIL	VANN	CRES	MART	ROLL	DECA	TECH	GORV	
PJ(J)	86.	44.	30.	125.	74.	37.	40.	44.	45.	23.	35.	22.	605.
chan	1.	0.	4.	5.	0.	0.	0.	2.	0.	0.	0.	1.	17.
prog	0.	0.	0.	0.	0.	1.	1.	0.	1.	2.	0.	1.	8.
droi	2.	1.	1.	8.	0.	1.	0.	0.	0.	0.	0.	0.	19.
libe	2.	10.	2.	5.	0.	1.	9.	0.	3.	4.	0.	2.	38.
inde	1.	0.	1.	0.	0.	4.	0.	0.	1.	0.	0.	2.	9.
demo	6.	13.	3.	3.	0.	0.	5.	1.	0.	3.	3.	8.	45.
dict	0.	0.	0.	0.	0.	0.	0.	0.	0.	3.	4.	0.	7.
trav	11.	0.	2.	28.	20.	7.	1.	3.	2.	1.	4.	0.	77.
fem	0.	1.	1.	6.	6.	0.	0.	1.	2.	0.	1.	0.	18.
sala	1.	2.	1.	2.	2.	0.	0.	0.	5.	0.	0.	0.	13.
prix	1.	2.	2.	0.	2.	0.	0.	1.	1.	1.	0.	0.	10.
choo	7.	2.	0.	2.	1.	0.	0.	1.	4.	1.	2.	1.	21.
empl	2.	1.	2.	1.	0.	0.	3.	1.	2.	0.	0.	1.	13.
oite	5.	0.	0.	6.	6.	11.	0.	2.	0.	0.	5.	0.	34.
gauc	6.	0.	0.	5.	9.	3.	0.	1.	1.	2.	5.	0.	32.
femi	0.	7.	1.	0.	0.	6.	1.	0.	4.	0.	0.	2.	21.
cont	3.	0.	1.	6.	7.	1.	2.	0.	0.	0.	0.	0.	20.
cri	7.	0.	0.	4.	1.	0.	2.	2.	0.	0.	4.	1.	21.
soci	8.	1.	8.	10.	6.	0.	5.	3.	2.	1.	4.	0.	48.
comm	5.	0.	0.	12.	4.	1.	6.	0.	4.	2.	2.	0.	36.
auto	0.	0.	0.	3.	0.	0.	1.	12.	0.	1.	0.	0.	17.
lutt	11.	2.	0.	8.	7.	1.	0.	3.	0.	0.	0.	1.	33.
smig	0.	2.	1.	2.	0.	0.	0.	1.	2.	1.	0.	0.	9.
nucl	0.	0.	0.	4.	1.	0.	0.	2.	0.	0.	0.	1.	8.
regi	0.	0.	0.	0.	0.	0.	2.	1.	1.	0.	0.	1.	5.
capi	7.	0.	0.	7.	2.	0.	0.	1.	1.	0.	1.	0.	19.
prco	0.	0.	0.	3.	0.	0.	2.	0.	3.	1.	0.	0.	9.

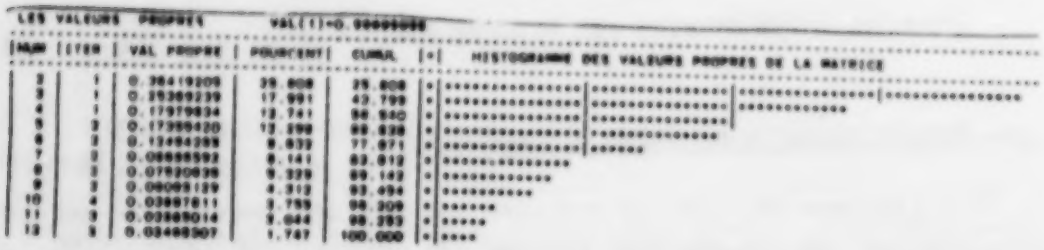
Le tableau de données

Les 27 mots sont en ligne. La dernière colonne contient les effectifs totaux (par exemple changement a été utilisé 17 fois).

Les 12 candidats sont en colonne. La première ligne contient les effectifs totaux (Ainsi LAFO a utilisé 86 fois les mots sélectionnés).

L'effectif du tableau apparaît en haut à droite : 605.

12- LES INERTIES PROJETEES



LES VALEURS PROPRES VAL(1)=0.

NUM	ITER	VAL PROPRE	POURCENT	CUMUL	HISTOGRAMME DES VALEURS PROPRES DE LA MATRICE
2	1	0.38419205	25.608	25.608	
3	1	0.29369239	17.991	43.799	
4	1	0.17978834	12.741	56.540	
5	2	0.17359420	13.598	69.838	
6	2	0.13484288	8.833	77.671	
7	2	0.08888592	6.141	83.812	
8	2	0.07920838	5.329	89.142	
9	2	0.06098129	4.312	93.454	
10	4	0.03897811	2.755	96.209	
11	3	0.02885014	1.844	98.253	
12	5	0.03488307	1.747	100.000	

La colonne Val Propre indique l'inertie de chacun des axes.
Les inerties sont les mêmes dans les deux nuages.

La colonne Pourcent indique le rapport entre l'inertie du nuage projeté sur chaque axe et l'inertie du nuage dans l'espace.

La colonne Cumul indique le rapport entre l'inertie du nuage projeté sur l'ensemble des premiers axes et l'inertie du nuage dans l'espace. C'est une mesure de la qualité de l'approximation que donnent les premiers facteurs. La projection sur le plan 1-2 représente par exemple 43,799% de l'inertie du nuage.

13- PROJECTION DES MOTS

	II	OLT	POID	INR	1AF	COR	CTR	2AF	COR	CTR	3AF	COR	CTR	4AF	COR	CTR	5AF	COR	CTR
1	chan	788	38	48	-388	66	11	499*	111	38	389	80	31	783	259	94	-809	399	147
2	prag	989	10	28	-971	340	36	-388	21	3	-312	35	9	-847	107	36	-830	176	98
3	drol	844	31	32	419	122	19	939	910	109	-389	95	34	117	9	3	109	8	3
4	libe	809	82	63	-1023	737	181	11	0	0	171	30	10	-171	31	11	158	37	19
5	inde	772	15	43	-435	48	8	-758	140	33	-1348	445	191	888	119	42	-297	33	11
6	demo	937	74	88	-1018	833	213	-49	1	1	-193	14	10	-232	67	47	980	213	208
7	diol	980	13	78	-140	3	1	-883	49	19	-28	0	0	-2810	719	494	-1354	194	170
8	trav	891	127	42	993	733	133	-224	109	25	92	8	2	137	29	14	98	7	3
9	fame	488	30	17	607	304	14	-21	1	0	340	142	19	298	107	18	-107	14	3
10	sala	790	21	27	-430	99	10	-48	1	0	929	197	33	788	332	71	-991	171	53
11	pris	800	17	19	-409	131	6	176	24	2	150	18	2	199	13	2	-138	14	2
12	chos	109	25	19	-190	30	2	-80	9	1	181	43	6	-111	18	3	-74	7	2
13	sept	811	21	17	-893	380	35	331	98	9	308	83	11	238	90	7	14	0	0
14	otto	944	98	85	872	199	50	-650	259	94	-988	480	338	43	1	1	-207	38	38
15	gaus	875	53	37	948	434	44	-450	287	42	-67	3	1	-310	137	29	-130	24	7
16	fael	933	35	88	-1081	421	107	-484	85	32	-938	393	138	848	187	84	-92	3	2
17	cont	898	33	30	529	329	35	-397	78	9	321	130	30	231	84	10	384	98	21
18	cri	903	35	33	309	103	9	20	0	0	66	4	1	-841	918	98	371	79	20
19	anci	308	79	37	100	21	2	108	25	4	350	299	94	-38	2	0	-19	1	0
20	comm	440	80	31	91	18	1	-98	18	3	431	369	81	-88	16	3	-111	25	6
21	auto	979	28	132	488	38	18	2182	776	538	-815	137	131	-382	20	20	-218	8	11
22	tutt	741	99	35	452	230	31	9	0	0	37	2	0	89	12	2	910	407	114
23	anig	829	19	14	-838	307	17	402	122	9	188	31	3	187	26	3	-449	193	34
24	nucl	934	13	15	414	108	6	759	354	30	-338	33	4	93	5	1	250	38	7
25	regi	370	9	21	-895	223	18	736	148	17	-13	0	0	4	0	0	-47	1	0
26	capi	891	31	18	987	378	27	0	0	0	313	119	17	-3	0	0	283	99	20
27	proc	847	19	23	-470	106	9	181	12	2	702	334	61	141	9	2	-179	288	73
			808.0	1000			1000			1000			1000			1000			1000

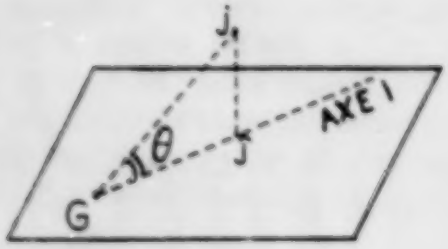
Les colonnes 1 * F, 2 * F etc.. contiennent les valeurs des facteurs : les coordonnées des projections des points sur chacun des axes. Ce sont ces valeurs qui sont utilisées pour les représentations graphiques planes.

(Pour les autres colonnes voir le paragraphe suivant).

14- PROJECTION DU NUAGE DES CANDIDATS

	JI	QLT	POID	INR	1**	COR	CTR	2**	COR	CTR	3**	COR	CTR	4**	COR	CTR	5**	COR	CTR
1	LAFO	487	142	52	293	169	33	-160	50	15	143	39	16	-83	14	6	338	219	130
2	LECL	811	73	116	-1241	686	308	-72	2	2	-181	15	13	88	2	2	487	106	139
3	HERV	364	50	56	-440	123	27	204	28	8	313	62	27	435	119	94	-231	34	22
4	PORT	877	207	91	399	480	90	110	35	10	203	131	52	107	33	14	79	18	10
5	GUIL	747	122	77	663	527	197	-381	140	60	188	39	24	159	29	18	102	12	10
6	VANN	954	81	118	155	9	4	-842	285	171	-1241	578	525	436	71	67	-298	33	43
7	CRES	474	68	74	-714	323	93	184	17	7	384	84	49	-222	32	19	167	18	15
8	MART	992	73	153	427	81	36	1525	784	666	-641	138	167	-115	5	6	-88	3	5
9	POLL	887	74	89	-985	189	88	54	2	1	405	97	88	818	228	184	-794	373	377
10	DECA	831	38	74	-628	144	41	-85	3	1	99	4	2	-1200	528	316	-849	154	129
11	TECH	824	88	78	323	57	17	-811	144	60	-95	5	3	-1001	551	335	-349	67	57
12	GORV	678	36	88	-1134	488	129	-78	2	1	-820	103	55	-48	1	0	466	82	63
			606.0	1000			1000			1000			1000			1000			1000

La colonne COR contient la qualité de représentation de chaque candidat sur chacun des axes. C'est (en millièmes) le cosinus carré de l'angle entre l'axe et la direction de point.



Projection de j sur l'axe 1

COR 1(j) = $\cos^2$ θ.

VANN, par exemple est très mal représenté par l'axe 1. Le cosinus carré de l'angle qu'il fait avec cet axe avait 0.009, il est donc presque orthogonal. Ce point se projette très près de l'origine G sur l'axe, sa distance à G qui est grande n'est pas visible sur l'axe.

La colonne CTR contient la contribution de chaque point à l'inertie de l'axe. Sur l'axe 2, MART à lui seul représente les 2/3 de l'inertie du nuage sur cet axe: il détermine cet axe.

15- LE PREMIER PLAN

Le graphique (cf graphe 1) montre la projection sur le plan déterminé par les axes 1 (horizontal) et 2 (vertical).

Nous interprétons chacun des facteurs indépendamment. Sur le premier facteur on trouve d'abord à gauche, les deux candidats de la droite traditionnelle qui dans leurs déclarations se réfèrent explicitement à des valeurs chrétiennes. Ils utilisent beaucoup les mots liberté, famille, démocratie, progrès et très peu les mots situés à l'opposé : capitalisme, travailleurs, gauche, contre, lutte, femme etc... Ces derniers sont au contraire très utilisés par les candidats situés à droite de l'axe. On trouve là tous les candidats des divers partis d'extrême gauche; on trouve aussi de ce côté de l'origine, le candidat d'extrême droite. Ce dernier est proche de l'origine, (sur l'axe horizontal), et nous l'avons vu mal représenté sur cet axe. Il est cependant beaucoup plus proche sur cet axe des candidats d'extrême gauche que des 2 candidats de droite. Le premier axe ne répartit donc pas les candidats suivant la classification droite-gauche.

A gauche de l'axe, mais beaucoup moins extrêmes que les deux premiers, on trouve les 4 grands partis : les deux partis de la majorité de l'époque, le parti socialiste, le parti communiste. Leurs sujets de préoccupation étant l'emploi, le salaire, les prix, le SMIG; leur proximité s'explique bien par le fait que ces 4 partis, seuls susceptibles d'avoir des candidats élus discutaient entre eux des mêmes sujets d'économie quotidienne. Les autres candidats privilégient plus les mots d'idées ou de principes : la famille pour les uns, lutte-capitalisme-travailleur pour les autres.

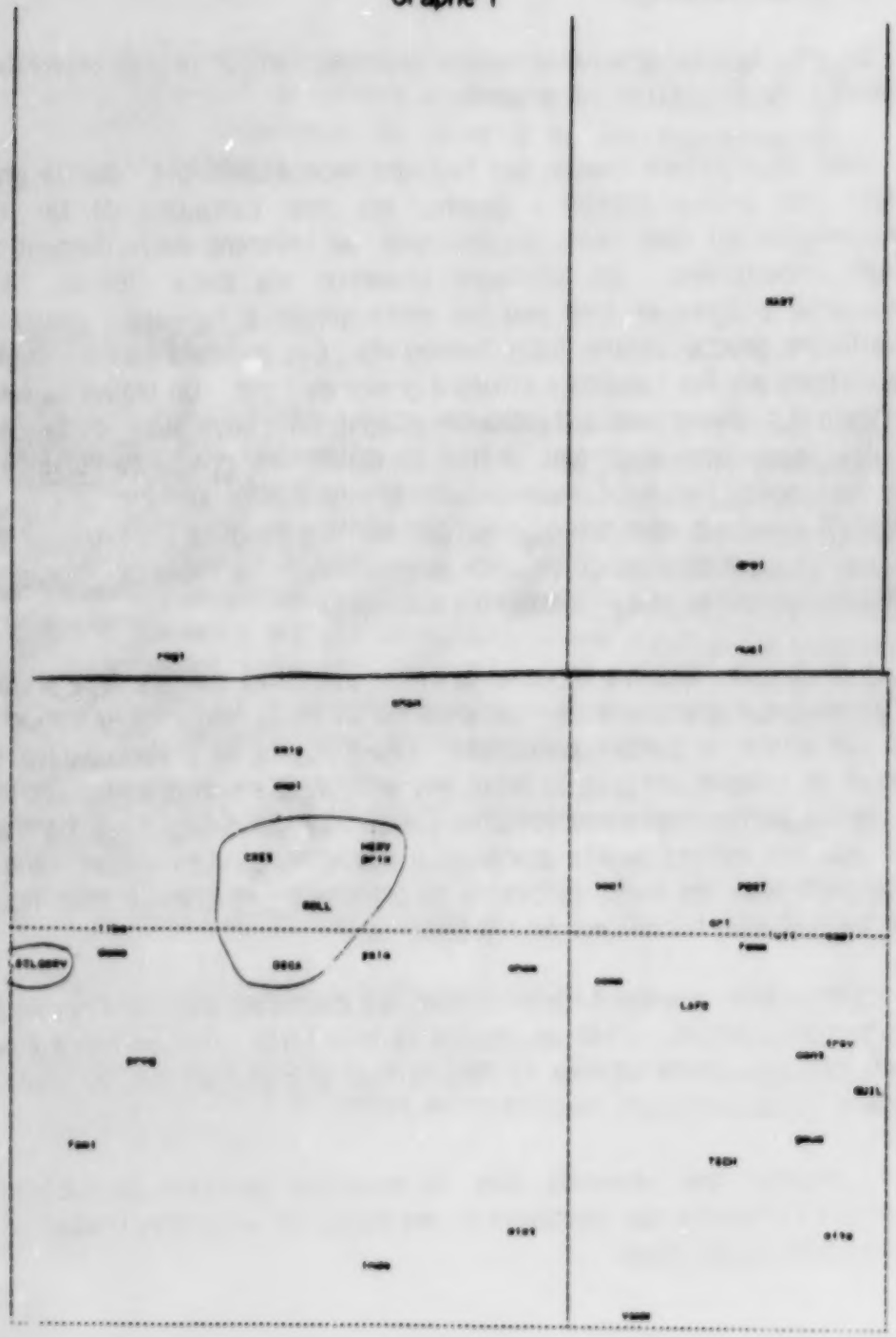
Le deuxième axe est déterminé pour les candidats par MART, et pour les mots par autogestion. C'est en réalité la très forte liaison entre eux que traduit cet axe : MART utilise 12 fois le mot autogestion qui, au total est utilisé 17 fois.

Le premier axe résumait donc la structure générale du tableau qui concernait l'ensemble des candidats et des mots. Le deuxième, traduit un lien ponctuel très prédominant.

PROJECTION DES 27 MOTS ET DES 12 CANDIDATS SUR LE PLAN 1-2

Axe horizontal (1)- axe vertical (2)--Titre: vocabulaire des députés.
Largeur = 1.92447 Hauteur = 3.02573 - Nombre de points = 39.

Graphe 1



DEUX METHODES D'ANALYSE DE DONNEES DIDACTIQUES:

CLASSIFICATION IMPLICATIVE ET CLASSIFICATION HIERARCHIQUE

par Régis Gras, Institut de Recherche Mathématique de l'Université de Rennes et laboratoire de recherche du GRECO "Didactique et Acquisition des Connaissances Scientifiques".

S1. PROBLEMATIQUE

La statistique classique centre son étude sur les tests et sur l'estimation de paramètres. Elle **nécessite** de ce fait, en général, la **présomption et l'émission d'hypothèses** en vue de les valider ou de les réfuter, que ces hypothèses s'expriment en termes de valeurs possibles ou bien de lois a priori ou encore d'homogénéité de population. Par exemple, on formulera l'hypothèse que la moyenne d'une certaine variable aléatoire (v.a.) appartient à un intervalle donné ou bien que la v.a. suit une loi normale ou bien que tous les éléments d'une population possèdent, selon certains critères, les mêmes caractéristiques. A l'aide d'une expérience de tirage aléatoire, on acceptera ou on réfutera cette hypothèse en s'accordant un seuil d'erreur lié à ce choix.

Or la conjoncture en 1980 est bien différente de celle du siècle dernier et même de 1950:

1) grand nombre de données recueillies

2) énormes possibilités de traitement par ordinateur

3) difficulté en didactique, qui tente de s'ériger en nouvelle science, de procéder par émission, refus ou acceptation d'hypothèses sur les phénomènes en présence.

Les **nouvelles méthodes d'analyse des données** répondent en partie à ces contraintes conjoncturelles. Elles **évitent** généralement **d'imposer** au spécialiste **la formulation d'hypothèses,** et au contraire, favorisent celle-ci pour des études plus classiques ultérieures. De plus, elles **permettent,** en condensant l'information de données très nombreuses, **de dégager des structures** ou des **facteurs** latents qu'il eût été difficile de percevoir autrement. Elles s'accompagnent toujours d'une visualisation géométrique des résultats. Par exemple:

nuages de l'analyse factorielle

noyaux des nuées dynamiques

arbre de classification hiérarchique

Ces représentations facilitent l'analyse puis l'interprétation par la mise en évidence de structures de condensation de l'information. Elles s'accompagnent de l'explicitation de degrés de pertinence de ces structures par rapport aux critères utilisés dans la méthode de traitement: par exemple, l'ordre de grandeur des inerties restituées par les axes factoriels, les variances des noyaux, les statistiques locales et globales des partitions à des niveaux de hiérarchie.

Les données didactiques classiques s'expriment en termes de comportements de réponse à des questionnaires (réussite - échec, modalités non binaires, rang de réponse, opinions, etc...) ou simplement d'attitudes observées en situation. Généralement, ces données sont très nombreuses, disparates et quelquefois hétérogènes dans leur nature. Le didacticien cherche des structures explicatives et prospectives dans le champ des comportements et des attitudes, mais aussi bien dans le champ des individus qui les manifestent. Leur étude, s'appuyant sur les liaisons respectives des 2 champs, fait accéder à une

certaine connaissance du croisement de leur structure.

Nous étudierons deux telles analyses qui visent à organiser des variables selon leur subordination ou selon leur ressemblance (ou leur dissemblance). Dans les deux cas, le principe d'association repose sur la même idée: on considère l'association d'autant plus intense qu'elle est invraisemblablement grande par rapport à celle qui serait observable dans le cas d'absence de lien entre les variables. Un arbre traduit ensuite la structure de subordination ou celle de similarité. Je vais spécifier ces deux méthodes de traitement de données en m'appuyant sur des exemples vécus.

§2. CLASSIFICATION IMPLICATIVE

2.1. Position du problème

En sciences humaines, un psychologue, un ethnologue, un pédagogue sont vivement intéressés par les relations d'implication ou de sériation entre variables (faits et comportements). Ils recherchent, tout comme le didacticien, à introduire un ordre ou un préordre total, sinon partiel, entre les réalisations des variables, afin de disposer d'une structure qui permette d'affirmer, avec une certaine certitude, non seulement que telle observation est indépendante de telle autre, mais aussi qu'elle admet nécessairement des prédécesseurs et des successeurs privilégiés. Notre problème, dans le cas de données didactiques binaires ou binarisées va donc être maintenant de donner un sens à l'expression: "le comportement a entraîne le comportement b". Ce sens sera accompagné de sa propre limitation, c'est-à-dire d'un seuil statistique au-delà duquel l'assertion sera réfutable. Un graphe orienté, pondéré, transitif illustrera la structure multidimensionnelle définie par cette relation d'implication sur l'ensemble des comportements envisagés. On notera l'ambition de cette méthode par rapport à celle classique de recherche d'échelle unidimensionnelle. Le travail d'analyse et d'interprétation de la structure constituera la dernière étape du traitement.

2.2 Indices d'implication

a) J. Loevinger a présenté en 1947 l'indice suivant pour apprécier l'implication: $a \Rightarrow b$, ou ce qui est équivalent, sa propre négation c'est-à-dire: $a \wedge \overline{b}$. Notons:

E(a): le sous-ensemble de la population E ayant eu le comportement a

$E(\overline{b})$: le sous-ensemble de la population E ayant eu le comportement $\neg b$

$n = \text{card } E$; $n(a) = \text{card } E(a) = |E(a)|$; $n(\overline{b}) = |E(\overline{b})| = n - n(b)$;

$p(a) = \frac{n(a)}{n}$ fréquence du comportement a

$p(\overline{b}) = \frac{n(\overline{b})}{n}$ fréquence du comportement $\neg b$.

L'indice de J. Loevinger est alors:

$$L(a,\overline{b}) = 1 - \frac{p(a \wedge \overline{b})}{p(a)p(b)} = \frac{p(a)p(\overline{b}) - p(a \wedge \overline{b})}{p(a)p(\overline{b})} = 1 - \frac{n \times n(a \wedge \overline{b})}{n(a)n(\overline{b})} \in]-\infty, 1].$$

C'est un indicateur d'implication qui prend la valeur 1 lorsque $E(a) \subset E(b)$ (implication stricte), la valeur 0 lorsque a et b sont indépendants. Mais certaines de ses valeurs peuvent être négatives et grandes en valeur absolue. Face à cette étendue de valeurs et devant le côté exceptionnel de l'inclusion stricte, il nous faut choisir un seuil d'acceptabilité de l'implication. Mais aucun argument fondé ne permet sérieusement de le faire. Que signifient les valeurs 0,9; 0,8; 0,5 par rapport à des fluctuations d'échantillonnage?

b) Inspiré par la démarche générale de Lerman dans la définition de la similarité entre deux variables, j'ai construit un <u>indice qui cherche à rendre compte de l'intensité d'implication</u>, tout en contrôlant par un test statistique ses limites d'admissibilité. Plusieurs approches sont possibles et nous en reparlerons lors de la construction de l'indice de similarité. Elles s'appuient toutes sur l'"invraisemblance" de la taille de $E(a \wedge \overline{b})$ eu égard à celles de E(a) et $E(\overline{b})$. Intuitivement, plus la taille de $E(a \wedge \overline{b})$ est petite par rapport à celle que l'on pourrait s'attendre à trouver dans le cas d'insubordination, plus forte est l'implication.

Deux éléments interviendront alors dans la forme générale de l'indice:

- d'une part, la valeur observée de $|E(a \wedge \bar{b})|$ dans l'expérience soit s
- d'autre part, la mesure en terme de probabilité de l'intervalle $]\leftarrow,s]$, i.e. Prob $[S \leq s]$.

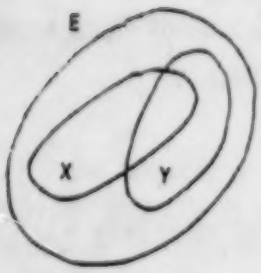
Dire que cette probabilité est faible revient à dire qu'il est improbable que la valeur trouvée soit aussi petite dans les conditions d'observations, sans qu'existe un phénomène de dépendance entre a et b. Par suite, l'implication conjecturée se trouve renforcée par la petitesse de Prob $[S \leq s]$.

c) Construction de l'indice

Une hypothèse d'absence de lien a priori entre a et b conduit à l'indépendance mutuelle, pour un élément quelconque de E, des événements:

"appartenir à E(a)" (avec la probabilité p(a)) et "appartenir à $E(\bar{b})$" (avec la probabilité $p(\bar{b})$)

Bien entendu, seules les situations $p(a) \leq p(b)$, sont, pour l'implication, intéressantes.



Notons X et Y des parties quelconques de E associées respectivement à E(a) et $E(\bar{b})$ telles que tout élément de E ait la probabilité p(a) (resp. $p(\bar{b})$) d'appartenir à X (resp. à Y). Sous l'hypothèse d'indépendance, l'appartenance simultanée à X et Y est $p(a)p(\bar{b})$.

Ainsi, la v.a. $|X \cap Y|$ est binomiale de paramètres n et $p = p(a)p(\bar{b})$

$$\text{Prob}\,[|X \cap Y| = s] = \binom{n}{s} [p(a)p(\bar{b})]^s [1-p(a)p(\bar{b})]^{n-s} \qquad (1)$$

Or, pour n assez grand et np assez petit, à cette v.a. on peut substituer une

v.a. de Poisson de paramètre $\lambda = np = n\,p(a)p(\bar{b})$. Par suite:

$$\text{Prob}[|X \cap Y| = s] = \frac{(np)^s \; e^{-np}}{s!} \qquad (2)$$

Cette 2ème approche est celle que nous retenons maintenant: elle est plus conforme à la volonté de "rareté" de $|X \cap Y|$. Par exemple, pour $n = 100$ $n(a) = 40$ et $n(\bar{b}) = 30$, alors: $np(a)p(\bar{b}) = 12$.

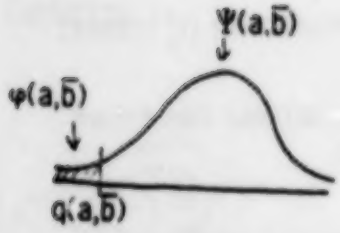
Enfin, un théorème de convergence en probabilité nous permet, dans la majorité de nos situations pratiques, de considérer que la v.a. $\frac{|X \cap Y| - \mu}{\sigma}$ suit approximativement une loi normale $N(0,1)$ avec $\mu = \sigma^2 = \lambda = np(a)p(\bar{b})$.

Dans notre démarche, il s'agit donc de placer la valeur observée $|E(a \wedge \bar{b})|$ au cours de l'expérience, dans le champ des valeurs possibles. Plus $|E(a \wedge \bar{b})|$ est petit dans ce champ, donc également $\frac{|E(a \wedge \bar{b})| - \mu}{\sigma}$, plus grande est l'hypothèse de: $a \longrightarrow b$.

Notons: $q(a,\bar{b}) = \frac{|E(a \wedge \bar{b})| - \mu}{\sigma} = \frac{n(a \wedge \bar{b}) - \frac{n(a)n(\bar{b})}{n}}{\sqrt{\frac{n(a)n(\bar{b})}{n}}}$ l'indice d'implication

et $\Psi[q(a,\bar{b})] = 1 - \varphi(a,\bar{b}) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q(a,\bar{b})} e^{-\frac{x^2}{2}}\,dx$ par référence à la loi normale $N(0,1)$.

D'où l'expression de notre test de décision pour $a \longrightarrow b$ qui ne fonctionnera, bien entendu, que dans le cas où l'implication est possible, à savoir: $n(a) \leq n(b)$.



La relation $a \longrightarrow b$ est admissible au seuil de $1 - \Psi_0$ ssi:
ou bien $E(a) \subset E(b)$
ou bien $\Psi[q(a,\bar{b})] \geq \Psi_0$

Dans la majorité des cas, notre "test d'implication" est défini au seuil de 5% (ici $\Psi_0 = 0{,}95$) comme dans les tests statistiques ordinaires.

Remarques

(1) Notons que $q(a,\bar{b})$ s'écrit encore:

$$q(a,\bar{b}) = \frac{np(a \wedge \bar{b}) - np(a)p(\bar{b})}{\sqrt{np(a)p(\bar{b})}} = \frac{\sqrt{n}\,p(a \wedge \bar{b}) - p(a)p(\bar{b})}{\sqrt{p(a)p(\bar{b})}}$$

On lui reconnait sous cette forme une affinité certaine avec l'indice de Loevinger $\dfrac{p(a)p(\bar{b}) - p(a \wedge \bar{b})}{p(a)p(\bar{b})}$.

(2) On aboutit à la formule (1) en considérant l'approche probabiliste suivante, par laquelle il s'agit de mesurer $\text{Prob}[|X \cap Y| = s]$. Pour cela, X et Y ont des cardinaux aléatoires, mais la probabilité qu'un élément quelconque de X (resp. Y) soit dans E(a) (resp. $E(\bar{b})$) est p(a) (resp. $p(\bar{b})$).

De plus, les appartenances à E(a) et $E(\bar{b})$ sont indépendantes. Si X et Y ont des cardinaux k et h respectivement, il s'agit dans l'épreuve aléatoire:

- de choisir k éléments parmi les n de E pour constituer $X(\binom{n}{k}$ choix)

- de choisir s éléments parmi les k pour constituer $X \cap Y$ ($\binom{k}{s}$ choix)
- de choisir h-s éléments parmi les n-k (ceux de $E \setminus X$) pour constituer $Y \setminus X$ ($\binom{n-k}{h-s}$ choix)
- d'imposer que les k éléments de X soient dans E(a) $(p(a)^k)$
- d'imposer que les n-k éléments de $E \setminus X$ soient dans $E(\bar{a})$ $(p(\bar{a})^{n-k})$
- d'imposer que les h éléments de Y soient dans $E(\bar{b})$ $(p(\bar{b})^h)$
- d'imposer que les n-h éléments de $E \setminus Y$ soient dans E(b) $((1-p(\bar{b}))^{n-h})$.
- de faire varier ensuite "convenablement" k et h:

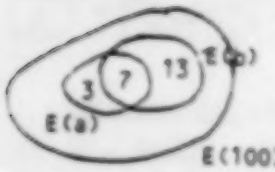
$$\text{Prob}[|X \cap Y| = s] = \sum_{\substack{(k,h) \\ k \geq s \\ h \geq s \\ k+h-s \leq n}} \binom{n}{k}\binom{k}{s}\binom{n-k}{h-s} [p(a)^k (1-p(a))^{n-k}] [p(\bar{b})^h (1-p(\bar{b}))^{n-h}].$$

En effectuant la sommation, on retrouve (1):

$$\text{Prob}[|X \cap Y| = s] = \binom{n}{s} p^s (1-p)^{n-s} \text{ où } p = p(a)p(\bar{b}).$$

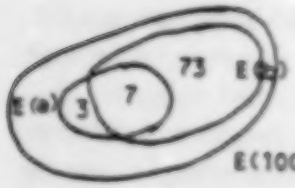
(3) Depuis peu, nous testons également l'implication à l'aide de l'indice $q(\bar{a},b)$, toujours associé à la condition initiale: $n(a) \leq n(b)$. Un calcul simple montre que: $\forall(a,b)\ q(a,\bar{b}) \leq q(\bar{a},b)$ ce qui, par la plus grande exigence de ce dernier indice, garantit son adéquation au problème. L'avantage de cet indice est de faire davantage ressortir les faibles variations de croissance de E(b) par rapport à E(a) supposé fixé. Si les comportements a et b sont des réussites à un questionnaire, $q(\bar{a},b)$ met en évidence les légères variations de complexité. Par contre, il présente une certaine inertie dans des situations où E(b) est très sensiblement plus "gros" que E(a).

Exemples



$$q(a,\bar{b}) = \frac{3 - \frac{10\times80}{100}}{\sqrt{\frac{10\times80}{100}}} = -1,77$$

$\Psi[q(a,\bar{b})]=0,96$ donc au seuil de 5%. $a \rightarrow b$



$$q(a,\bar{b}) = \frac{3 - \frac{10\times20}{100}}{\sqrt{\frac{10\times20}{100}}} = 0,71$$

$\Psi[q(a,\bar{b})]=0,24$ donc $\neg(a \rightarrow b)$ au même seuil

2.3 Construction d'un préordre partiel sur l'ensemble des comportements.

En calculant pour tous les couples (a,b) tels que $n(a) \leq n(b)$ l'indice $q(a,\bar{b})$, nous pouvons construire une matrice carrée d'éléments génériques $\eta(i,j)$ définis ainsi:

$\eta(i,j) = 1$ si $\Psi[q(i,\bar{j})] \geq \Psi_0$ ($= 0,95$ en général)
$= 0$ sinon.

On obtient une matrice dite d'implication de la forme:

↗	a	b	c	d	e	f
a	1	1	0	0	1	0
b	0	1	0	0	0	0
c	0	0	1	0	0	1
d	0	0	0	1	0	1
e	0	1	0	0	1	0
f	0	0	0	1	1	1

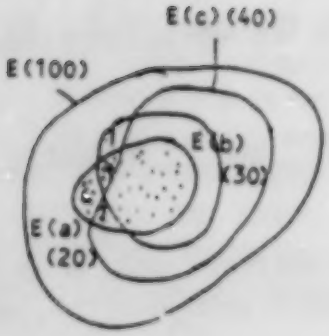
On remarque que:

$\eta(f,e) = 1$, $\eta(e,b) = 1$ et
la case $\eta(f,b)$ est vide.

L'implication stricte est transitive. Par suite, elle permet de construire un préordre partiel dans l'ensemble des variables. Ce n'est plus le cas dans la quasi-implication définie par le test ci-dessus. Nous pouvons simplement affirmer que si $a \rightarrow b$ et $b \rightarrow c$ au même seuil Ψ_0, il est probable que $a \rightarrow c$ à ce seuil. On a d'ailleurs l'inégalité triangulaire suivante:

$$q(a,\bar{c}) \leq \sqrt{\frac{n(\bar{b})}{n(\bar{c})}}\, q(a,\bar{b}) + \sqrt{\frac{n(b)}{n(a)}}\, q(b,\bar{c}) + \frac{1}{\sqrt{\mu(a,\bar{c})}} [\mu(a,\bar{b}) + \mu(b,\bar{c}) - \mu(a,\bar{c})]$$

$$\text{où } \mu(x,\bar{y}) = \frac{n(x)n(\bar{y})}{n}$$



Par exemple:

$n = 100$; $n(a) = 20$; $n(b) = 30$; $n(c) = 40$
$n(a \wedge \bar{b}) = 5$ $n(a \wedge \bar{c}) = 5$ $n(b \wedge \bar{c}) = 2$

on obtient:

$q(a,\bar{b}) = -2,41$ $q(b,\bar{c}) = -3,77$.

Au seuil de 5%, $a \rightarrow b$ et $b \rightarrow c$.

L'inégalité précédente montre que: $q(a,\bar{c}) \leq -1,45$, alors qu'en réalité $q(a,\bar{c}) = -2,02$, et donc montre que $a \rightarrow c$.

Afin d'obtenir le préordre partiel attendu sur l'ensemble des variables selon la relation de quasi-implication, nous relèverons si nécessaire le seuil d'admissibilité rendant ainsi la relation transitive dans des limites raisonnables (au pire au seuil $\Psi_0 = 0,50$).

2.4 Graphe d'implication

Nous pouvons maintenant construire le graphe orienté, transitif, sans boucle (sauf en cas de symétrie), pondéré (à l'aide des $\Psi[q(a,\bar{b})]$ ou des $\eta(a,\bar{b})$, associé à la relation de quasi-implication. Par construction, le graphe admet:

pour sources des variables qui ne sont successeurs d'aucune autre
pour puits des variables qui n'ont pas de successeurs.

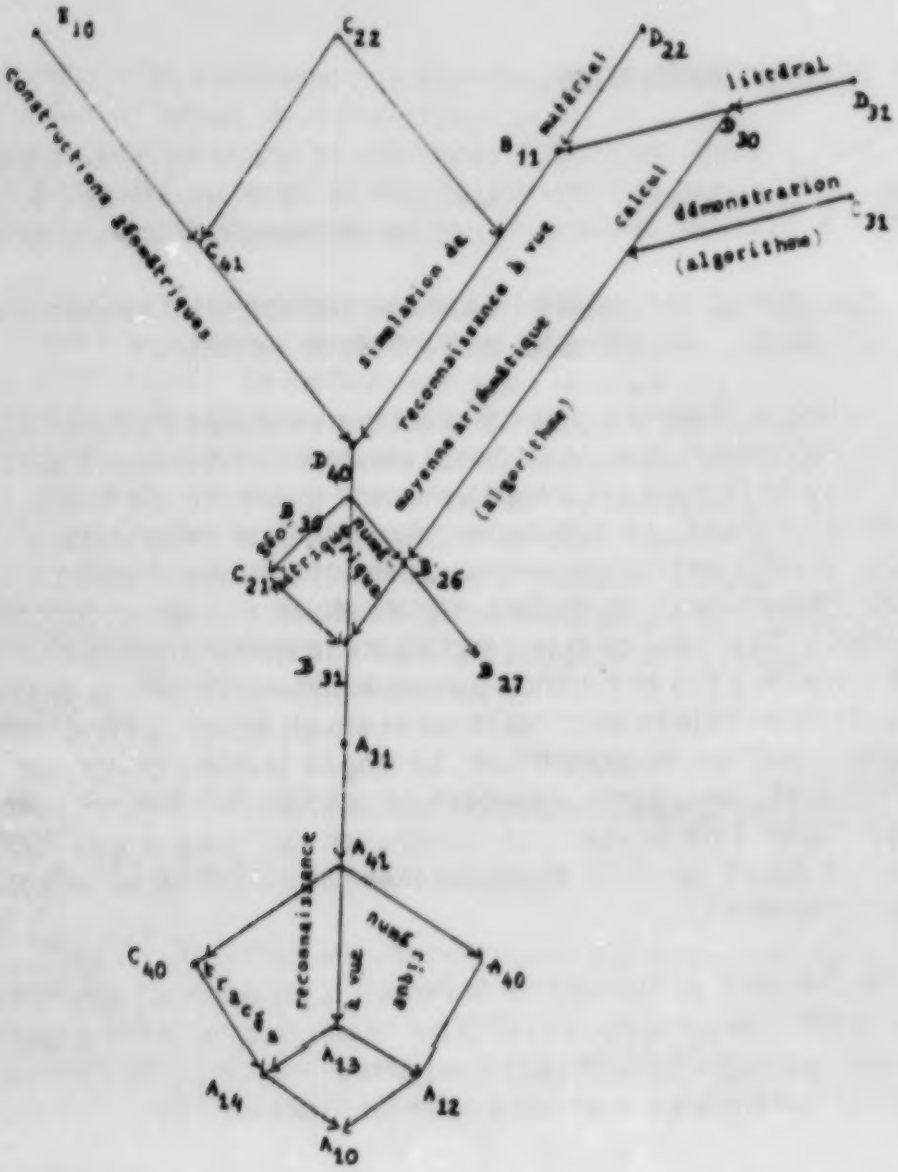
Le long de la même chaîne s'alignent des successeurs transitifs de la même source, successeurs dont les cardinaux associés vont croissant d'une source à un puit. Nous ne fermons pas transitivement le graphe (cf. exemple du test sur la symétrie centrale). Le didacticien intervient dans l'interprétation du graphe obtenu. Il porte particulièrement son attention sur l'enchaînement des variables et sur l'indépendance de chaînes, significatives de champs comportementaux différents. C'est ainsi qu'un test portant sur la symétrie centrale (cf. graphique page suivante) a mis en évidence, par une analyse de contenu et de tâche, que la discrimination majeure entre les items se situait autour de la représentation du concept, numérique ou géométrique. Le long de la même chaîne, par contre, se hiérarchisent les items sensiblement comme le laissait prévoir leur identification taxonomique. La conjonction des sous-arbres "contenus" se produit à nature de tâche égale (par exemple un calcul ou un dessin défini algorithmiquement).

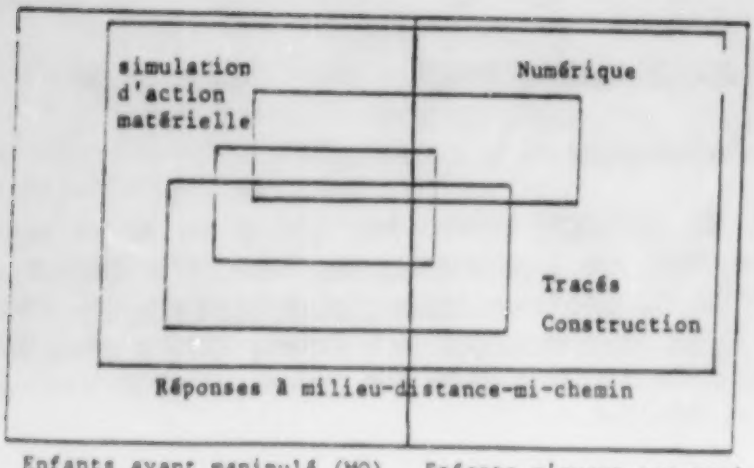
Un chercheur de l'Université de Rennes, H. Rostam s'est penché récemment dans sa thèse de 3ème cycle (1981) sur la construction automatique de l'arbre que nous ne savions jusqu'en 1981 que construire à la main. H. Ratsimba-Rajohn, de l'I.R.E.M. de Bordeaux, a amélioré cette construction.

§3- CLASSIFICATION HIERARCHIQUE

3.1 Position du problème

Cette fois, on recherche comme dans toutes les autres méthodes de classification automatique, à constituer, sur l'ensemble V des variables, des partitions de moins en moins fines, construites de façon ascendante à l'aide d'un critère de similarité entre variables. Les données, comme précédemment, se présentent généralement sous la forme d'un tableau: $V \times E$ où V représente les variables et E les individus.





Enfants ayant manipulé (MO) Enfants n'ayant pas manipulé (MN)

Le didacticien s'intéresse à ce type d'analyse qui lui permet d'étudier puis d'interpréter en termes de ressemblance (ou de dissemblance) des noyaux de variables, constitués significativement à certains niveaux de l'arbre et s'opposant à d'autres à ces mêmes niveaux. En même temps, il définit une typologie sur l'ensemble V des variables.

I.C. Lerman a permis de faire évoluer les méthodes de classification grâce à cette hypothèse d'absence de lien entre variables, réinvestie à la base de l'indice d'implication. Sa méthode est assez générale pour prendre en compte différents types de variables, mettant ainsi un outil d'analyse très puissant à la disposition de nombreux utilisateurs. Sans s'opposer aux méthodes factorielles, elle a sur elles l'avantage de présenter synthétiquement en un arbre, l'information condensée d'une structure ou d'une famille de structures sur l'ensemble des variables. Pour ma part, j'associe analyse factorielle des correspondances, analyse en classification hiérarchique et maintenant analyse implicative (totale ou partielle dans V) à chacun des systèmes de données fournies par des enquêtes en milieu scolaire.

3.2 Typologie des principales variables traitées par cette méthode

Une classification possible est la suivante:

- catégorie des variables représentables par une partie de E ou par une pondération sur E
- catégorie des variables représentables par une partie de $E \times E$ ou par une pondération sur $E \times E$.

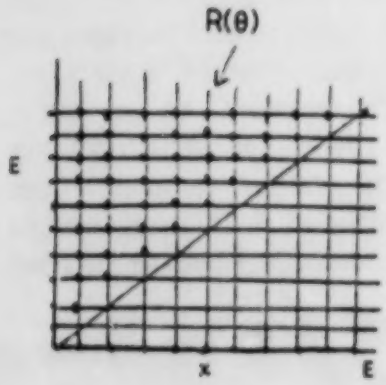
1ère catégorie

a) attribut descriptif: c'est le cas d'une question "a" d'un test où, à tout élève $x \in E$, on associe $a(x) = 1$ (resp. 0) s'il possède (resp. ne possède pas) l'attribut "a" ou, de même, si sa réponse à "a" est juste (resp. fausse). Ainsi, "a" est représentable par la partie $E(a) = a^{-1}(1)$.

b) variable numérique: par exemple, une question "a" d'un test étant donnée, a(x) est le nombre ou la proportion d'élèves de la classe scolaire x à avoir bien répondu à "a". Cette variable apparaît donc comme une pondération sur E: $x \xrightarrow{a} a(x)$.

2ème catégorie

a) variable rang: une note ou un classement indiquent une performance d'enfants à une certaine épreuve "a". On suppose ici que l'échelle de notes ou le classement définissent un ordre total θ sur E (pas d'ex-aequos).



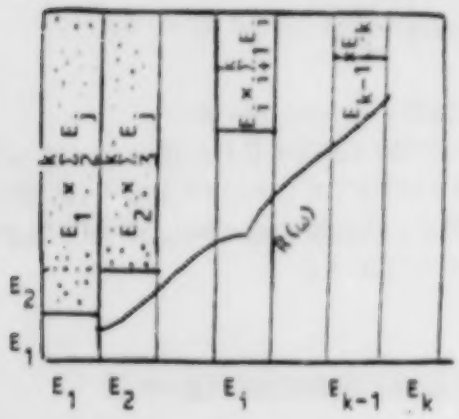
La variable rang associée est alors représentable par le graphe de θ dans $E \times E$: $R(\theta) = \{(x,y) \in E \times E \mid x < y \text{ pour } \theta\}$ où nous notons par exemple $x < y$ ssi $a(x) < a(y)$. Rappelons que: $\text{card } R(\theta) = \frac{n(n-1)}{2}$ où $\text{card } E = n$.

b) variables ordinales (caractères à modalités ordonnées).

C'est le cas des questions d'une enquête où les modalités sont totalement ordonnées, comme par exemple:

"beaucoup", "un peu", "pas du tout".

Au caractère "a", dont l'ensemble des modalités est $\{m_1, m_2, \ldots, m_k\}$, est associée l'application: $x \longmapsto a(x) = m_i$, si x a choisi la modalité m_i. Cette application définit un préordre total ω sur E, lui-même associé à la partition $E = \bigcup_{i=1}^{k} E_i$ où $E_i = a^{-1}(m_i)$. Le préordre ω se représente par la partie $R(\omega)$ de $E \times E$:



$$R(\omega) = \{(x,y) \mid x < y \text{ et } \neg(y < x)\}$$

$$= \sum_{i<j}^{\omega} E_i \times E_j \text{ (somme ensembliste).}$$

Remarquons, pour le calcul ultérieur de l'indice:

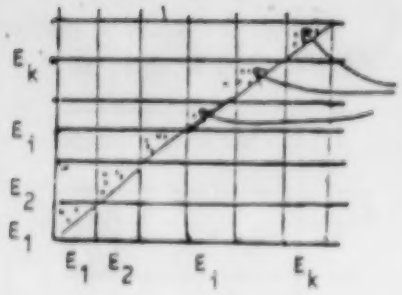
$$\text{card } R(\omega) = \sum_{i<j} n(i)\, n(j) \text{ où } n(i) = \text{card } E_i$$

$$n(j) = \text{card } E_j$$

$$= \sum_{i \leq k-1} n(i) \sum_{j<i} n(j).$$

c) <u>variables partition</u> (caractère sans structure de modalités).

De la situation précédente, il suffit d'exclure l'ordre entre les modalités, comme par exemple des stratégies employées par le maître pour introduire une notion "a": "manipulation", "dessin", "problème", etc.... A chaque variable "a", on peut faire correspondre une partition $\pi = \{E_1, \ldots, E_k\}$ où $E_i = a^{-1}(m_i)$. A π, on peut associer l'ensemble $R(\pi)$ des paires réunies par π, sous-ensemble de l'ensemble F de toutes les paires de E:



$$R(\pi) = \{\{x,y\} \in F \mid \exists i \in \{1,\ldots,k\}\ (x,y) \in E_i \times E_i\}$$

$$\text{card } R(\pi) = \sum_{i=1}^{k} \frac{n(i)(n(i)-1)}{2} \text{ où } n(i) = \text{card } E_i$$

On peut également représenter π par l'ensemble des paires séparées par π.

d) <u>variables pondération</u>: un ensemble d'items E étant donné, définissons une pondération p sur $E \times E$ à l'aide de l'indice d'implication précédent en posant.

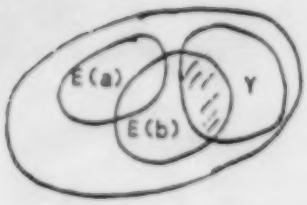
$\forall(a,b) \in E^2$: si $n(a) \leq n(b)$ $p(a,b) = \Psi[q(a,\bar{b})]$
si $n(a) > n(b)$ $p(a,b) = 0$.

Chaque variable apparait comme une matrice carrée $|E| \times |E|$. p est une variable pondération dont on peut chercher la similarité avec une autre variable p' définie ainsi: $\forall(a,b) \in E^2$: si b appartient à une classe taxonomique inférieure à celle de a, $p'(a,b) = 1$. Dans le cas contraire, $p'(a,b) = 0$.

3.3 Indice de similarité entre deux attributs descriptifs

La démarche initiale de I.C. Lerman a été la suivante:

Soient a et b deux attributs, E(a) et E(b) les représentations de ces attributs sur E, n(a) et n(b) les cardinaux respectifs de ces classes. Le cardinal s de $E(a) \cap E(b) = E(a \wedge b)$ est un indicateur de la similarité de a et b, biaisé par les tailles de E(a) et E(b). Aussi, associant à E(b) une partie aléatoire Y choisie au hasard dans le sous-ensemble de $\mathbb{P}(E)$ des parties de même cardinal n(b), Lerman considère la v.a. S_a = card [E(a) ∩ Y] sous une hypothèse d'absence de lien a priori entre E(a) et Y. Comme nous l'avons fait dans le §2, la similarité entre a et b est d'autant plus grande que la valeur observée $|E(a) \cap E(b)|$ est grande par rapport à celles qui sont attendues.



L'indice de similarité initial sera donc: Prob $[S_a \leq s]$.

Remarquons que: $\text{Prob}[S_a = k] = \dfrac{\binom{n(a)}{k}\binom{n-n(a)}{n(b)-k}}{\binom{n}{n(b)}}$

La loi de S_a est hypergéométrique:

de moyenne $\mu = \frac{n(a)\,n(b)}{n}$

de variance $\sigma^2 = \frac{n(a)\,n(\bar{a})\,n(b)\,n(\bar{b})}{n^2(n-1)}$

Si nous choisissons plutôt X parmi les parties de cardinal n(a), la v.a. $S_b = |X \cap E(b)|$ suit la même loi que S_a. On montre de plus que, pour n assez grand, la v.a.

$S(a,b) = \frac{S_a - \mu}{\sigma}$ ou $\frac{S_b - \mu}{\sigma}$ suit une loi normale N(0,1). Cette approximation s'avère très utile pour les calculs dans la majorité des cas. En général:

$P(a,b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{s(a,b)} e^{-\frac{x^2}{2}}\,dx$ est choisi comme indice de proximité entre a et b, où s(a,b) est la valeur observée de S(a,b).

Le carré de s(a,b) est le χ^2 associé au tableau de contingence:

	E(b)	$E(\bar{b})$	
E(a)	$s=n(a\wedge b)$	u	n(a)
$E(\bar{a})$	v	t	$n(\bar{a})$
	n(b)	$n(\bar{b})$	

Observations

	E(b)	$E(\bar{b})$	
E(a)	$\frac{n(a)n(b)}{n}$	$\frac{n(a)n(\bar{b})}{n}$	n(a)
$E(\bar{a})$	$\frac{n(\bar{a})n(b)}{n}$	$\frac{n(\bar{a})n(\bar{b})}{n}$	$n(\bar{a})$
	n(b)	$n(\bar{b})$	

Fréquences théoriques

On notera ici le changement de point de vue: le χ^2 ne se place pas par

rapport à un seuil χ^2_0 d'indépendance, mais sert de mesure de similarité.

L'indice s(a,b) s'écrit encore:

$$s(a,b) = \sqrt{n}\,\frac{st - uv}{\sqrt{(s+u)(s+v)(t+u)(t+v)}}$$

La forme symétrique de s(a,b) en s et t donne autant de poids, dans la similarité, aux associations positives qu'aux associations négatives. Pour cette raison, nous employons à la place de cet indice hypergéométrique l'indice binomial et plutôt l'indice poissonnien définis comme précédemment par

indice binomial:
$$s_b(a,b) = \frac{n(a\wedge b) - \frac{n(a)n(b)}{n}}{\sqrt{\frac{n(a)n(b)}{n}\left(1 - \frac{n(a)n(b)}{n^2}\right)}} < s(a,b)$$

indice poissonnien:
$$s_p(a,b) = \frac{n(a\wedge b) - \frac{n(a)n(b)}{n}}{\sqrt{\frac{n(a)n(b)}{n}}} < s_b(a,b) < s(a,b).$$

Ce dernier indice valorise, pour $p(a) \cdot p(b) < 1$, les associations positives entre a et b, ce qui, pour caractériser la proximité entre a et b, événements rares, est préférable aux yeux du didacticien.

Remarque. Les indices de proximité entre variables numériques et pondération, variables rang, variables ordinales et partition, s'inspirent de la définition de l'indice hypergéométrique entre variables-attribut. Signalons la construction spécifique des variables numériques.

Si $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ et $(\beta_1, \beta_2, \ldots, \beta_n)$ sont les vecteurs indicateurs des attributs a et b sur E (α_i = 1 ou 0 suivant que l'individu i possède ou

non l'attribut a), remarquons alors que: $s = |E(a \wedge b)| = \sum_{i=1}^{n} \alpha_i \beta_i$.

De la même façon, pour les variables numériques a et b, considérons:

$$S_a = \sum_i \alpha_i \beta_{\tau(i)} \text{ et } S_b = \sum_i \alpha_{\sigma(i)} \beta_i \text{ où } \alpha_i = a(x_i) \text{ et } \beta_i = b(x_i).$$

où τ et σ sont des permutations aléatoires, éléments de l'ensemble équiréparti des permutations de $\{1,2,\ldots,n\}$, telles que:

$$\sum_i \beta_{\tau(i)} = \sum_i \beta_i \text{ et } \sum_i \alpha_{\sigma(i)} = \sum_i \alpha_i.$$

Les 2 indices S_a et S_b suivent la même loi de moyenne μ et d'écart-type σ. $S(a,b) = \frac{Sa-\mu}{\sigma}$ suit approximativement la loi $N(0,1)$.

3.4 Indice de proximité entre classes

Dans une première étape de l'algorithme de construction de l'arbre de classification, on calcule l'indice P(a,b) pour toutes les paires de variables. Mais la classification doit se poursuivre sur un nouvel ensemble constitué de paires et de singletons, puis sur un ensemble de classes d'éléments d'effectif quelconque.

I.C. Lerman retient alors l'indice de base suivant, homologue de $s = |E(a \wedge b)|$ dans le cas des singletons.

Soient 2 parties disjointes A et B de V:

$$p(A,B) = \max_{(a,b) \in A \times B} [P(a,b)] = \max_{a \in A} [\max_{b \in B} P(a,b)]$$

Or l'ensemble des nombres $\pi(a,b)$, dont $P(a,b)$ est une réalisation, constitue, dans l'hypothèse d'absence de lien, lorsque b décrit B de cardinal m et lorsque a est fixé dans A, un échantillon de m points indépendants, réalisations d'une v.a. uniformément répartie sur $]0,1[$. Par suite:

$$\forall t \in]0,1[: \text{ Prob } [\pi(\{a\},B) < t] = \text{Prob } [\forall b \in B, \pi(a,b) < t] = t^m.$$

De même, l'ensemble des nombres $\pi(\{a\},B)$, lorsque "a" décrit A de cardinal ℓ, est un échantillon de ℓ points indépendants. Par suite:

$$\text{Prob } [\pi(A,B) < t^m] = \text{Prob } [\forall a \in A, \pi(\{a\},B) < t^m] = (t^m)^{\ell}.$$

On retiendra donc comme indice de proximité entre A et B, en faisant $t = p(A,B)$, valeur observée:

$$P(A,B) = \text{Prob } [\max_{(a,b)\in A\times B} \pi(a,b) < p(A,B)] = [p(A,B)]^{\ell m}.$$

3.5 Arbre de classification

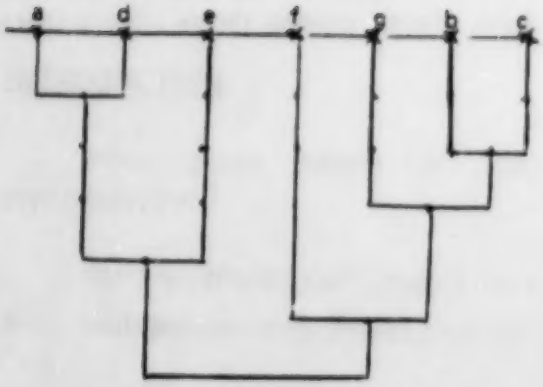
L'algorithme de la construction de l'arbre est donc le suivant:

- On part de la partition la plus fine sur V. On agrège en une classe les éléments a et b les plus proches, c'est-à-dire ceux pour lesquels P(a,b) est maximum;

- à partir de la partition obtenue (en général, il reste (n-2) singletons et une paire), on calcule tous les indices de proximité par le procédé P(A,B) ci-dessus. On agrège les 2 classes (1 paire et 1 singleton ou 2 singletons) les plus proches pour cet indice;

• on continue en effectuant ainsi en tout, en général, card V - 1 condensations jusqu'à l'obtention de la partition de V réduite à V lui-même.(1)

Exemple.



{{a},{d},{e},{f},{g},{b},{c}}
{{a,d},{e},{f},{g},{b},{c}}
{{a,d},{e},{f},{g},{b,c}}
{{a,d},{e},{f},{g,b,c}}
{{a,d,e},{f},{g,b,c}}
{{a,d,e},{f,g,b,c}}
{{a,d,e,f,g,b,c}}.

Une statistique locale permet de mettre en évidence la constitution significative de classes à certains niveaux où la statistique connaît un maximum. Cette connaissance intéresse l'analyste: la partition obtenue à ce niveau est susceptible d'interprétation, car une classe vient de s'y former de façon consistante. Un arbre condensé, réduit à ces niveaux significatifs est fourni avec l'arbre complet de classification et c'est celui que nous analysons généralement.

---

(1) Les programmes informatiques de classification et d'analyse implicative fonctionnent sous Multics. J'ai élaboré des logiciels de ces traitements qui, pour des nombres de variables et individus "raisonnables", sont opérationnels en Basic sur T07-70.

Un travail du laboratoire de I.C. Lerman (1979) permet de croiser toute partition intéressante, fournie par l'arbre et la statistique locale, partition dite classification floue, et une autre classification floue. Mais également, il est possible de croiser une classification floue et une classification dite nette, constituée par exemple à partir de paramètres descriptifs de la population soumise à un questionnaire. Ce croisement que nous utilisons assez systématiquement permet de connaître quel type de paramètre intervient particulièrement dans la définition d'une classe donnée de la classification floue.

METHODES EXPLORATOIRES ET CONFIRMATOIRES EN ANALYSE DE LA REGRESSION

Communication présentée par Richard Bertrand, Université Laval, Faculté des Sciences de l'éducation, département de mesures et évaluation.

INTRODUCTION

"Nous avons besoin de l'analyse exploratoire et de l'analyse confirmatoire".[1]

Voilà un énoncé fort louable surtout qu'il provient d'un statisticien qui avait souligné quelques années plus tôt :

"Pendant longtemps, j'ai cru que j'étais un statisticien qui s'intéressait aux inférences (donc à l'analyse confirmatoire) allant du particulier au général. Mais attentif au développement de la statistique mathématique, j'ai trouvé des raisons d'étonnement et de doute".[2]

C'est sous l'impulsion de ce chercheur, John Tukey, que l'analyse exploratoire ou plutôt l'exploration en analyse statistique des données a été reconnue comme importante en soi.

Depuis, plusieurs ont aussi fait leur marque dans ce domaine : Frédérick Mosteller, David Hoaglin, Paul Velleman, John Emerson.

---

(1) "We need both exploratory and confirmatory", par John Tukey, The American Statistician, Vol 34, No 1, 1980.

(2) Traduction d'un texte de John Tukey dans "The Future of data analysis", Annals of Mathematical Statistics, Vol. 33, 1962, par G. Morlat dans la préface au volume de F. Cailliez et J.P. Pages, Introduction à l'analyse des données, SMASH, Paris, 1976.

Cependant, une des critiques que l'on peut formuler encore aujourd'hui à l'égard de l'analyse exploratoire est son manque de liens concrets avec l'analyse confirmatoire. La terminologie introduite par Tukey est souvent déroutante parce que distincte de celle couramment utilisée en statistique et on a quelquefois l'impression que l'emphase de l'analyse exploratoire est trop exclusivement descriptive. Ce qui semble nuire à sa diffusion.

Il m'apparaît que nous pouvons tirer beaucoup plus de l'exploration si on la perçoit comme une phase de l'analyse statistique des données reliée possiblement à une autre phase, confirmatoire ou inférentielle, mais bien plus encore si on s'imprégne d'une attitude exploratoire dans tous nos travaux d'analyse. Il me semble que c'est ce qu'ont fait les grands chercheurs expérimentaux de tout temps, avec ou sans ordinateur.

Au cours de cet exposé, je rappellerai en premier lieu les principales distinctions entre l'analyse exploratoire et l'analyse confirmatoire. Puis, je soulignerai les concepts et les méthodes les plus utilisés en analyse exploratoire. Je présenterai enfin une utilisation de l'approche exploratoire en analyse de la régression.

ANALYSE EXPLORATOIRE ET ANALYSE CONFIRMATOIRE

Classiquement, on distingue la statistique descriptive de la statistique inférentielle suivant que l'on veut décrire certaines caractéristiques des données ou que l'on veut faire des inférences, c'est-à-dire estimer des paramètres ou tester des hypothèses. Cependant, il m'apparaît que cette distinction ne rend pas suffisamment l'aspect dynamique d'une analyse statistique.

C'est pourquoi je préfère parler, comme Tukey, d'analyse exploratoire et d'analyse confirmatoire. Ce qui nous amène à distinguer deux phases lors d'une analyse statistique. Une première phase vise à explorer, c'est-à-dire non seulement décrire les caractéristiques des données mais aussi regarder intensément les relations entre les données, indiquer les particularités d'ensemble des données, détecter les valeurs aberrantes. Une seconde phase vise à confirmer, c'est-à-dire estimer des paramètres, tester des hypothèses. La première phase est nécessaire à la seconde et la prépare. Ce qui ne signifie pas qu'une analyse statistique n'est constituée que de deux phases. Il existe

une interaction entre exploration et confirmation qui montre qu'une analyse statistique doit être vue comme une procédure itérative où se succèdent phases exploratoires et phases confirmatoires.

Le tableau I présente une synthèse des principales distinctions entre l'approche exploratoire et l'approche confirmatoire.

Exploratoire	Confirmatoire
. Approche descriptive	. Approche inférentielle
. Logique simple	. Logique rigide
. Indication des hypothèses à tester	. Test d'hypothèse
. Utilisation d'indices statistiques résistants	. Utilisation d'indices moins résistants.
. Plan de recherche souple	. Plan de recherche rigoureux
. Usage des données en mains	. Usage de données exemptes d'erreur (si possible)
. Importance de l'expression visuelle ou graphique des données	. Moindre importance de l'expression visuelle ou graphique
. Avant une phase confirmatoire	. Après une phase exploratoire
. Vision intuitive des données	. Vision précise des données
. Esprit ouvert, divergent	. Esprit fermé, convergent
. Analogie avec une enquête policière	. Analogie avec un procès
. Importance du jugement du chercheur	. Importance du jugement du consultant
. Moindre importance de l'ordinateur	. Grande importance de l'ordinateur

Tableau I Principales distinctions entre A.E.D. et A.C.D.

CONCEPTS DE L'ANALYSE EXPLORATOIRE

Une analyse exploratoire se fait généralement à partir de statistiques résistantes comme la médiane, l'étendue interquartile et aussi à partir de graphiques simples et "signifiants" pour mieux décrire les données de façon suggestive et aussi pour bien préparer, le cas échéant, une phase confirmatoire.

Prenons, par exemple, le diagramme en feuilles représenté au graphe 1.

		(Nb)
4	1 5 6 7 7 8 8 9	(8)
5	0 1 6 7 7 9	(6)
6	1 2 3 4 5 6 7	(7)
7	4 6	(2)
		(n = 23)

Graphe 1. Diagramme en feuilles de 23 données.

Il représente la distribution des 23 données 41, 45, 46,---74, 76.

Chaque chiffre de la colonne située à gauche du trait vertical est appelé une branche (ou tige): il représente le premier chiffre composant une donnée. Chaque chiffre situé à droite du trait vertical est une feuille: il représente le second chiffre composant une donnée. Ainsi la donnée 49 est représentée par le chiffre (la feuille) 9 situé à l'extrême droite de la ligne débutant par le chiffre (la branche) 4.

Les données sont donc gardées à vue tout en permettant une vision graphique de la distribution de fréquences.

Parce que les données sont en ordre, il est relativement facile de localiser la médiane et les quartiles dans ce diagramme. La médiane (md) est la donnée de rang $\frac{1 + n}{2}$ où "n" est le nombre de données. Ici n = 23,

alors le rang de la médiane est 12 et donc md = 57.

Les quartiles inférieurs (Q_1) et supérieur (Q_3) ont rang

$$\frac{1 + \text{rang (md)}}{2}$$

Lorsque les données sont placées en ordre croissant (resp. décroissant) on obtient Q_1 (resp. Q_3).

Ici, le rang de ces quartiles est $\frac{1 + 12}{2} = 6,5$

Ainsi $Q_1 = \frac{48 + 48}{2} = 48$

et $Q_3 = \frac{63 + 64}{2} = 63,5$

Le diagramme en feuilles est donc fort utile pour représenter globalement les données, la forme de la distribution et localiser la médiane et les quartiles. Le graphe 2 donne un exemple d'un diagramme en feuilles plus complexe.

			(Nb)	(Pr)
	0*	3 4 4 4 5 5 6 9	(8)	(8)
	1	2 4 4 8	(4)	(12)
	2	1 4 7 9 9	(5)	(17)
	3	4 6	(2)	(19)
	4	0 2 9	(3)	(22)
	5	6 6	(2)	(24)
Q_1 →	6	1↓ 2 3 4 9	(5)	(29)
	7	0 8 8	(3)	(32)
	8	3 9	(2)	(34)
	9*	0	(1)	(35)
	1**	0↓ 1 3 3 4 4 4 6 6 7 7 8 8 9	(14)	(49)
Md →	2	2↓ 2 2 5 5 5 5 5 5 5 5 7 9	(13)	----
Q_3 →	3	2 2 2 4 5 5 6 6 7 9 9 9 9 9↓	(13)	(38)
	4	0 3 4 5 6 7	(6)	(25)
	5	2 2 2 3 3 4 4 7	(8)	(19)
	6	4 6 6	(3)	(11)
	7	5	(1)	(8)
	8	2 3 5	(3)	(7)
	9**	2 7	(2)	(4)
	1***	1 9	(2)	(2)
	2			

(T = 100)

Branche X $10^{Nb\,(*)}$; feuille X $10^{Nb\,(*)-1}$

Graphe 2. Diagramme en feuilles du nombre d'étudiants (par 100 000 habitants) engagés dans un programme d'enseignement supérieur pour 100 pays.

Cependant le diagramme en feuilles a ses limites. Il est difficile, par exemple, de bien visualiser l'étendue interquartile (EI). De la même façon, il peut garder cachées des valeurs aberrantes ("outliers"). De plus, il peut difficilement rendre compte de façon adéquate de la comparaison entre plusieurs distributions.

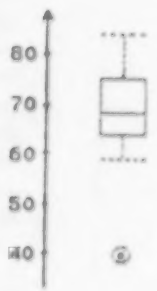
Prenons l'exemple suivant :

		(Nb)
4	0	(1)
5	9 9	(2)
6	4 4 8 8 9	(5)
7	5 5 7	(3)
8	1 3	(2)
		(n=13)

Graphe 3. Diagramme en feuilles recelant une valeur aberrante.

La distribution a l'air globalement correcte. Cependant, en y regardant de plus près, la donnée 40 semble se démarquer sensiblement des autres: elle est située à plus de 2 EI de Q_1. Pourtant le diagramme en feuilles ne rend pas bien cette évidence.

Le diagramme en boîte est présenté afin de combler ce genre de lacune. Ce diagramme, construit à l'échelle, se présente de la façon suivante:

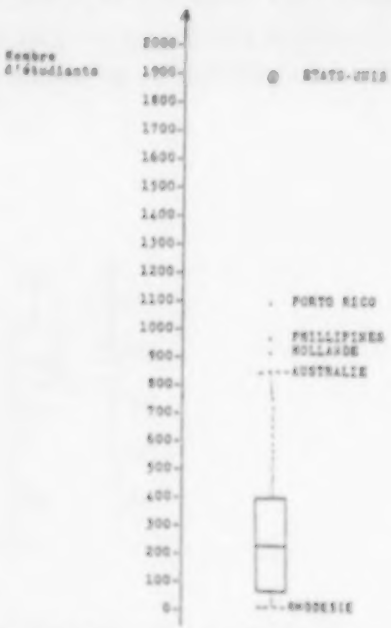


Graphe 4. Diagramme en boîte des données du graphe 3.

On commence par tracer l'échelle. Puis on place un trait horizontal vis-à-vis les valeurs de la médiane (md), du quartile inférieur (Q_1) et du quartile supérieur (Q_3). On relie ensuite ces trois traits horizontaux par deux barres verticales, de manière à former une boîte (un rectangle). On trace des traits pointillés horizontaux en haut et en bas de cette boîte aux endroits correspondant respectivement à la valeur la plus élevée des données mais ne dépassant pas Q_3 + 1 1/2 EI et à la valeur la moins élevée des données mais pas plus basse que Q_1 - 1 1/2 EI. Les données situées au-delà de Q_3 + 1 1/2 EI et en deça de Q_1 - 1 1/2 EI sont entourées et considérées comme potentiellement aberrantes.

Ainsi, ce diagramme permet de visualiser d'un seul coup d'oeil la médiane, les quartiles inférieur et supérieur, l'étendue interquartile, les valeurs potentiellement aberrantes, en plus de donner une idée assez précise de la symétrie de la distribution et de permettre la comparaison de plusieurs échantillons.

Le graphe 5 présente un diagramme en boîte plus élaboré.



Graphe 5 Diagramme en boîte du nombre d'étudiants (par 100 000 habitants) engagés dans un programme d'enseignement supérieur pour 100 pays.

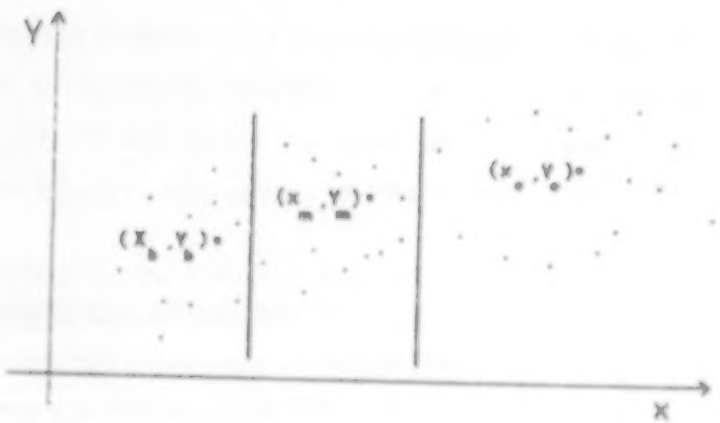
UTILISATION DE L'ANALYSE EXPLORATOIRE EN REGRESSION

On a vu que l'analyse exploratoire permettait de décrire, de visualiser les données de manière à effectuer une étude adéquate des diverses caractéristiques (symétrie, nombre de sommets, détection de valeurs aberrantes) d'une distribution de données. Cette étude peut se révéler fort utile avant d'entreprendre une phase confirmatoire.

L'analyse exploratoire offre également des outils très pertinents pour l'étude de la relation entre deux variables, à savoir l'étude d'un nuage de points.

Voici donc une méthode de construction d'une droite de régression résistante (appelée "droite de Tukey") basée sur le calcul de points sommaires résistants aux valeurs aberrantes.

La première opération consiste à diviser les données (points (X,Y)) en trois groupes de tailles semblables. Ce qui donne un groupe de valeurs basses, un groupe de valeurs situées au milieu et un groupe de valeurs élevées tel que l'on peut le constater au graphe 6. Plusieurs règles très strictes doivent être suivies afin d'arriver à une division satisfaisante (voir en particulier Velleman & Hoaglin [1981, p. 123]).



Graphe 6. Division du nuage de points en trois groupes et points sommaires.

Puis on calcule les points sommaires résistants pour les trois groupes (X_b, Y_b), (X_m, Y_m), (X_e, Y_e). La plupart du temps, on utilisera la médiane sur X et la médiane sur Y pour les points sommaires. Quelquefois, cependant, la moyenne tronquée ("trimmed mean", voir Mosteller & Tukey "1977, p. 34 ") ou la trimoyenne ("trimean", voir Tukey "1977, p. 46 ") est utilisée. La moyenne tronquée, notée Mt, est obtenue en enlevant la portion (environ 10%) des données située à chaque extrémité de la distribution et en calculant la moyenne arthmétique du reste des données. La trimoyenne, notée TM, est la moyenne arithmétique pondérée de la médiane et des quartiles inférieur et supérieur en donnant deux fois plus d'importance à la médiane.

Ainsi $$Mt = \sum_{I}^{(0,8)n} \frac{Xi}{(0,8)n} \quad \text{et} \quad TM = \frac{Q_1 + 2md + Q_3}{4}$$

L'étape suivante consiste à calculer, par approximations successives, la pente b_f de la droite cherchée. La première approximation est donnée par

$$b_1 = \frac{Ye - Yb}{Xe - Xb}$$

Les premiers résidus partiels sont les $Y_I - b_1 X_I$. La médiane de ces résidus pour le groupe des valeurs basses est notée Y'_b et la médiane des résidus pour les groupes des valeurs élevées est notée Y'_e. La première pente résiduelle vaut

$$b'_1 = \frac{Y'e - Y'b}{X_e - X_b}$$

Si $|b'_1| \leq 0{,}0001 \ |b_1|$, on retient b_1 comme pente finale : c'est-à-dire $b_f = b_1$.

Si $|b'_1| > 0{,}0001 \ |b_1|$, on prend comme <u>deuxième approximation</u> $b_2 = b_1 + cb_1$, où "c" est un nombre naturel : c= 1,2,3_ _ _ _ _ .

On veut b_2 tel que b'_2 soit de signe contraire à b'_1, de manière à bien encadrer la pente cherchée. Habituellement c=1 suffit pour que $\text{sgn}(b'_2) = \text{sgn}(b'_1)$; sinon, il faut prendre tour à tour c=2, puis c=3 et ainsi de suite.

Une fois b_2 trouvée, on calcule les deuxièmes résidus partiels $Y_i - b_2 X_i$ puis les médianes de ces résidus Y''_b et Y''_e et la pente résiduelle

$$b'_2 = \frac{Y''_e - Y''_b}{X_e - X_b}$$

Si $|b'_2| \leq 0{,}0001 \; |b_1|$, on retient b_2 comme pente finale : c'est-à-dire $b_f = b_2$.

Si $|b'_2| > 0{,}0001 \; |b_1|$, on prend comme troisième approximation,

$$b_3 = b_2 - b'_2 \left[\frac{(b_2 - b_1)}{(b'_2 - b'_1)} \right]$$

Puis on calcule la pente résiduelle b'_3 et on procède comme ci-haut.

En général, pour la "x" ième approximation,

$$b_x = b_s - b'_s \left[\frac{(b_s - b_t)}{(b'_s - b'_t)} \right]$$

où $x \geq 3$, b_s est la pente qui a la plus petite pente résiduelle b'_s (en valeur absolue) parmi celles qui ont déjà été calculées et b_t est la pente dont la pente résiduelle b'_t est la plus petite (en valeur absolue) parmi celles qui sont de signe contraire à b'_s.

Le principe directeur consiste donc en l'encadrement de la pente cherchée b_f par approximations successives. On calcule enfin l'ordonnée à l'origine a_f qui est la médiane des derniers résidus partiels $Y_i - b_f X_i$ prise sur l'ensemble des données.

La droite de Tukey est alors $\hat{Y} = a_f + b_f X$.

La construction de cette droite peut servir plusieurs fins. Nous traiterons de l'étude de la curvilinéarité et de la détection des valeurs aberrantes.

ETUDE DE LA CURVILINEARITE

Cette étude est basée sur l'alignement des trois points sommaires. Plus ces points sont alignés, plus l'ajustement linéaire risque d'être adéquat.

On calcule en premier lieu les pentes

$$b_{bm} = \frac{Y_m - Y_b}{X_m - X_b} \quad , \quad b_{me} = \frac{Y_e - Y_m}{X_e - X_m}$$

puis le rapport r_p de ces deux pentes en plaçant au numérateur la plus petite des pentes (en valeur absolue).

Si $r_p \geq 0,9$, l'ajustement linéaire est jugé adéquat.

Si $0,5 \leq r_p < 0,9$, il peut être utile de transformer les données, au jugement de l'analyste.

Si $0 \leq r_p < 0,5$, il vaut mieux effectuer une transformation monotone des données.

Si $r_p < 0$, ce peut être l'indice d'un effet non monotone (quadratique, cubique ou autre). On s'en rendra compte en effectuant des tests de tendance curvilinéaire (régression polynomiale).

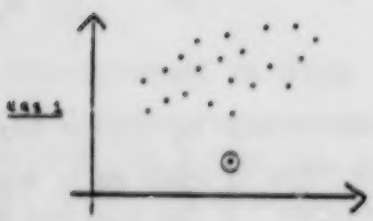
Dans le cas où l'on juge qu'il vaut mieux effectuer une transformation monotone des données, il reste à décider du type de transformation et de la variable à transformer.

La transformation, qui prend la forme $Z = T^q$ où q est un nombre rationnel, dépend de l'allure de la courbe et de l'intensité de la courbure.

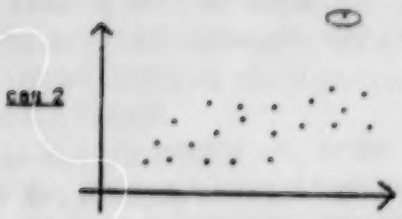
La variable à transformer dépend de la forme des distributions en cause. En général, on transforme plutôt la variable qui s'écarte le plus d'une forme standard (symétrique, un sommet, pas de valeur aberrante) de manière à obtenir une forme similaire pour les distributions de X et de Y. Quelquefois il est nécessaire de transformer les deux variables. L'étude préalable des diagrammes en feuilles et des diagrammes en boîte peut aider.

Détection des valeurs aberrantes

Deux cas classiques de valeurs aberrantes peuvent survenir :



Graphe 7. Valeur aberrante marginale



Graphe 8. Valeur aberrante dominante.

Au graphe 7, toutes les données sont typiques, c'est-à-dire bien regroupées ensemble, sauf une dite **marginale**. On détecte les valeurs marginales en examinant les résidus standardisés bien connus Zx. On peut parler de valeur marginale lorsque $|Zx| \geq 2$.

Au graphe 8, toutes les données sont typiques sauf une dite **dominante**.

On détecte les valeurs dominantes en calculant l'indice

$$h_x = \frac{(X - m_x)^2}{v_x\,(n-1)} + \frac{1}{n}$$

où m_x est la moyenne des n données x,

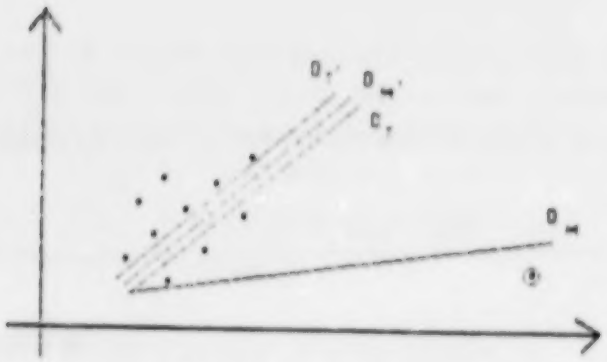
et v_x est la variance des n données x,

qui mesure le degré d'écart de l'abscisse X de chaque donnée (X,Y) à la moyenne m_x.

On a $\frac{1}{n} \leq h_x \leq 1$ et $\sum_{1}^{n} h_x = 2$.

Plus une donnée a une valeur h_x élevée plus cette donnée risque d'influencer de façon substantielle la pente de la droite de régression; c'est pourquoi elle est dite dominante. Il est possible d'évaluer si une donnée dominante (généralement si $hx \geq \frac{2}{n}$) a une influence véritable en employant la technique du "jackknife": c'est-à-dire recalculer la pente de la droite de régression en excluant, à tour de rôle, les données jugées dominantes puis comparer la pente de cette nouvelle droite avec la pente de la droite originale.

En procédant de la sorte, on s'aperçoit que la droite de régression (classique) des moindres carrés est peu résistante aux valeurs dominantes. Il est intéressant, dans ce contexte, de comparer son comportement à celui de la droite de Tukey. Examinons les quatre droites de régression du graphe 9. D_T est la droite de Tukey et D_M la droite des moindres carrés : elles sont calculées à partir de toutes les données.



Graphe 9. Droite des moindres carrés (DM)
Droite de Tukey (DT)

$D_{T'}$ et $D_{M'}$ sont respectivement la droite de Tukey et la droite des moindres carrés mais calculées sans tenir compte de la valeur dominante ⊙. Comme on peut le constater, les pentes des deux droites de Tukey étant semblables alors que les pentes des deux droites des moindres carrés sont passablement distinctes, la droite de Tukey est plus résistante aux valeurs dominantes que la droite des moindres carrés.

Je voudrais vous présenter maintenant un exemple provenant de données réelles dans le cadre d'une étude internationale de l'IEA. L'intérêt de cette étude était la recherche des relations entre les pratiques pédagogiques des enseignants et le rendement scolaire de leurs élèves. Vingt-et-une classes québécoises participaient à l'étude.

Prenons la pratique pédagogique qui consiste pour le professeur à approuver, féliciter le groupe dans un contexte d'enseignement collectif. Une classe obtient une valeur hx = 0,85 alors que le hx moyen pour 21 classes est d'environ 0,1. Ainsi cette classe a une valeur hx huit fois plus grande que la moyenne.

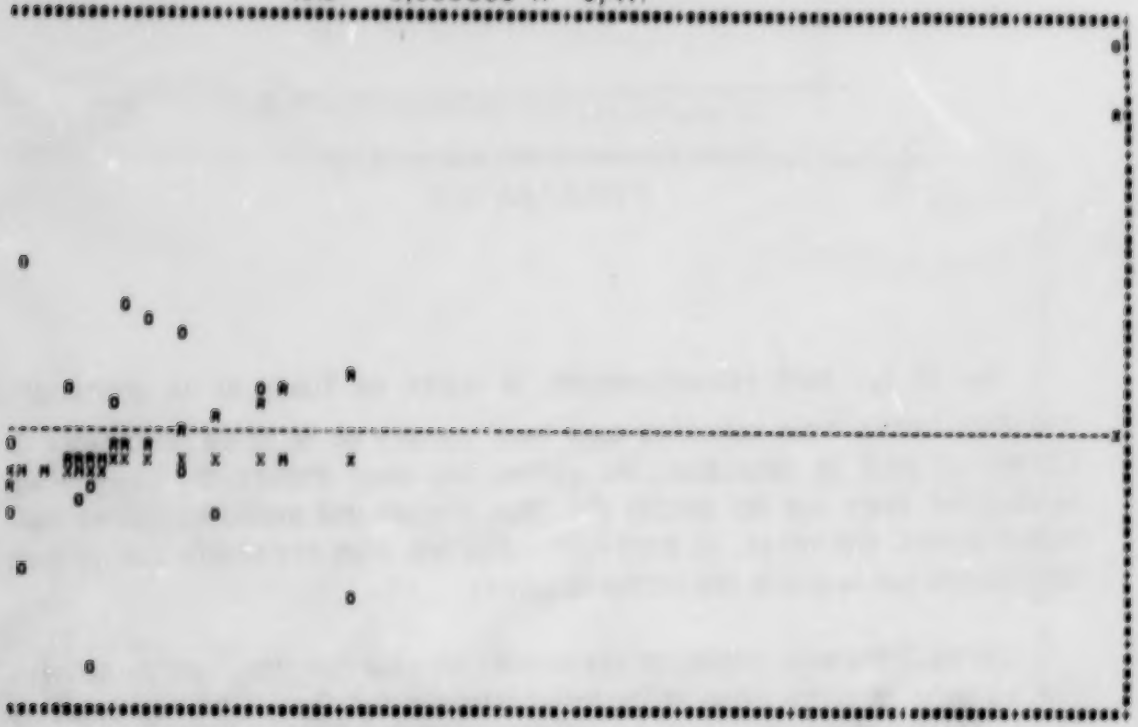
On a affaire ici à une valeur dominante.

Il faut donc porter une attention tout à fait particulière à cette donnée. Voyons un peu le nuage de points (graphe 10): il est facile de détecter la classe

délinquante!

En calculant la droite de régression des moindres carrés, on obtient dans ce cas-ci :

$$Y_{RMC} = 0{,}033805\ X - 0{,}41.$$



Graphe 10. Nuage de points (0), droite de Tukey (X) et droite des moindres carrés (R) représentant la relation entre les félicitations du groupe et le rendement scolaire.

La droite de régression de Tukey donne

$$Y_{RT} = 0{,}003961\ X - 0{,}28\ .$$

Le coefficient de corrélation est $r = 0{,}60$

On peut voir le tracé de ces droites au graphe 10 (points "R": moindres carrés, points "X": Tukey).

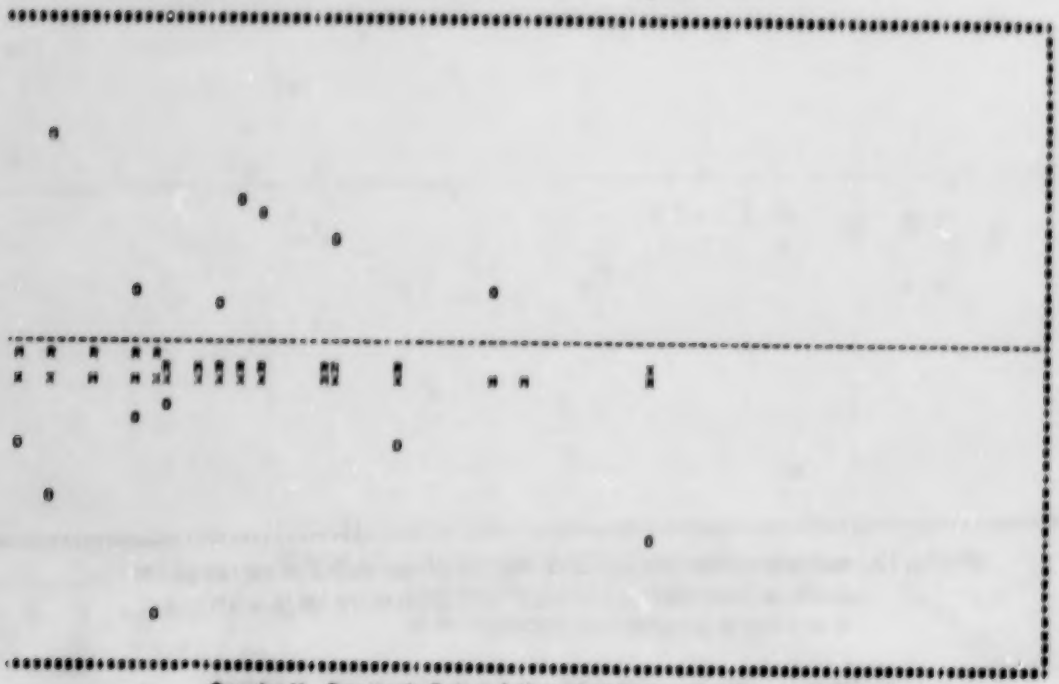
Les pentes de ces deux droites sont bien distinctes : l'une a une pente près de dix fois plus grande que l'autre. Si on enlève (par "jackknife") cette classe aberrante et que l'on recalcule les équations des droites, on obtient :

$$Y^{*}_{REC} = -0,006856\ X - 0,07$$

$$Y^{*}_{RT} = 0,003981\ X - 0,28$$

et $$r^{*} = -0,06.$$

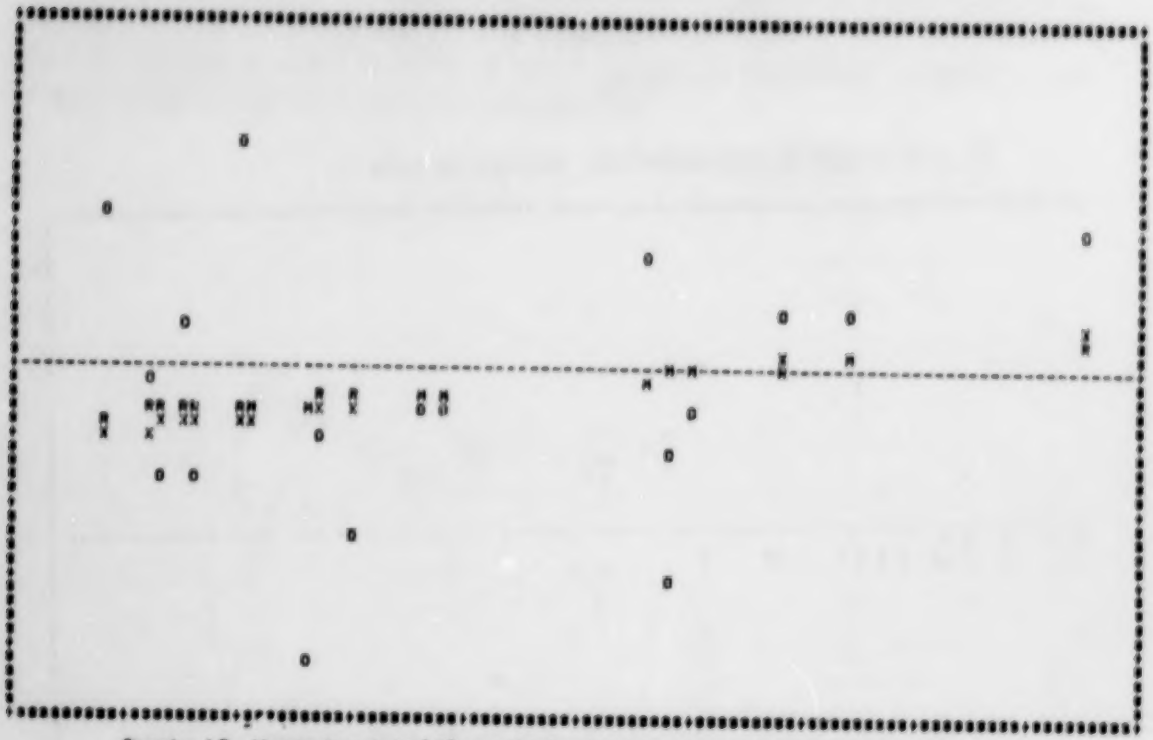
Voici le tracé de ces nouvelles droites (graphe II).



Graphe II. Droite de Tukey (X) et droite des moindres carrés (R) après excision de la classe aberrante du graphe 10.

Encore ici on constate que la droite de régression des moindres carrés est moins résistante que la droite de régression de Tukey. Mais il y a plus : le coefficient de corrélation s'annule (ou presque) si on enlève la classe aberrante. Ainsi, il n'est pas possible d'induire une relation linéaire forte et positive entre les félicitations de l'enseignant au groupe et le rendement scolaire, même si cela semblait clair (r=0,60) avant le jackknife.

Regardons maintenant le graphe 12, représentant la relation entre le rendement scolaire et la pratique pédagogique qui consiste à féliciter un élève en particulier.



Graphe 12. Nuage de points (0), droite de régression des moindres carrés (R) et droite de Tukey(X) représentant la relation entre les félicitations à un élève et le rendement scolaire.

On a $Y_{RMC} = 0,005\,X - 0,34$

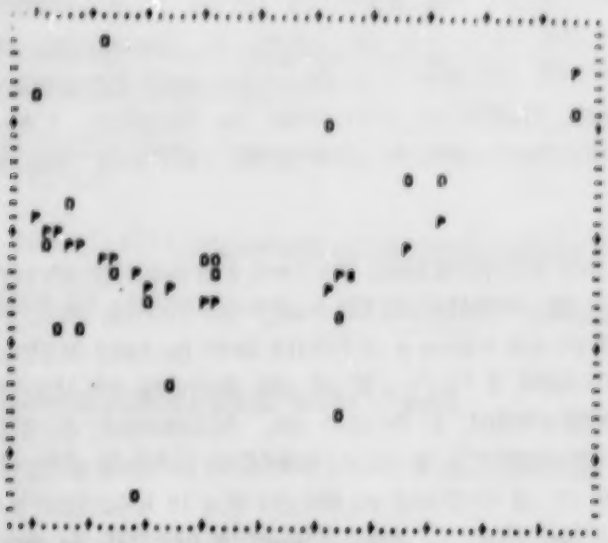
$Y_{RT} = 0,007\,X - 0,45$

et $r = 0,17$.

Comme on peut le voir sur le graphe 12 et à l'examen des équations, les deux droites de régression ont des pentes similaires. Cependant le lien linéaire semble assez faible puisque le coefficient de détermination est $r^2 = 0,03$ ($r = 0,17$) .

Mais comme le rapport des pentes $b_{bm} = -0,012$ et $b_{me} = 0,025$ est négatif, alors on peut soupçonner un effet curvilinéaire non monotone.

Effectivement, si on fait un test de tendance curvilinéaire, on obtient r^2 = 0,20 lorsque l'on ajoute l'effet quadratique. On peut le voir en regardant au graphe 13 le tracé de la courbe prédite dans le nuage de points. Ainsi, bien que l'on n'ait pas trouvé un effet linéaire important, il existe un effet quadratique qui montre que le rendement diminue avec une augmentation des félicitations à l'élève jusqu'à un certain point à partir duquel le rendement augmente avec l'augmentation des félicitations individuelles.



Graphe 13. Courbe prédite (P) exprimant la relation quadratique entre les félicitations à un élève et le rendement scolaire.

CONCLUSION

La régression linéaire est un modèle mathématique très utilisé en analyse des données. Or, trop peu souvent se préoccupe-t-on d'étudier à fond les conditions d'application du modèle (normalité des distributions, homogénéité des variances, linéarité) lorsque l'on veut confirmer certaines hypothèses. La phase exploratoire préalable est bien souvent capitale pour détecter des inadéquations (ex. valeurs aberrantes, curvilinéarité) entre les données et le modèle. Les méthodes exploratoires telles le diagramme en feuilles, le diagramme en boîte, la droite de Tukey, le "jackknife", les valeurs hx, les transformations sont des exemples de ce qui peut être employé du moins à la manière de Tukey, Mosteller, Velleman ou Hoaglin. L'essentiel cependant, au-delà des techniques, reste bien plus l'attitude exploratoire qu'il faut conserver.

Tukey et ses acolytes nous invitent à travailler en véritables détectives devant un paquet de données ou un nuage de points, avec ou sans techniques sophistiquées, je dirais même à la limite avec ou sans ordinateur. Tukey et ses acolytes nous invitent à faire parler les données en les scrutant à fond et même en les transformant, si besoin est. Autrement, aurait-on jamais appris que le taux d'accroissement de la période d'un pendule simple est proportionnel à la racine carrée de sa longueur ou encore que le logarithme de la pression d'un gaz est en relation linéaire avec l'inverse négatif de sa température? On n'insistera jamais assez pour affirmer que les grands chercheurs comme Clapeyron, Clausius et Hertz étaient aussi et d'abord, à l'enseigne des Pearson, Fisher et Tukey, de grands analystes de données.

REFERENCES

BERTRAND, R. et C. VALIQUETTE : Pratique de l'analyse statistique des données. Presses de l'Université du Québec (sous presse).

ERICKSON, B.H. et T.A. NOSANCHUK (1977): Understanding data. McGraw-Hill Ryerson Ltd.

HARTWIG, F., et B.E. DEARING (1979): Exploratory data analysis SAGE Publications Inc.

HOAGLIN, D.C., MOSTELLER, F., TUKEY, J.W., (1983): Understanding robust and exploratory data analysis, John Wiley & Sons.

LEINHARDT, G., et S. LEINHARDT, (1980) Exploratory data analysis : New tools for the analysis of empirical data, Review of Research in Education, Vol. 8.

MOSTELLER, F. et J. W. TUKEY (1968): Data analysis, including statistics The Handbook of Social Psychology, Vol. II, pp. 112-122.

MOSTELLER, F. et J.W. TUKEY (1977): Data analysis and regression, Addison-Wesley.

MOSTELLER, F., FIENBERG, S.E., ROURKE, R.E.K., (1983): Beginning Statistics with data Analysis, Addison-Wesley.

TUKEY, J. W. (1977): Exploratory data analysis, Addison-Wesley.

VELLEMAN, F., HOAGLIN, D.C., (1981): Applications, Basics and Computing of Exploratory Data Analysis, Wadsworth, Inc.

WAINER, H., et D. THISSEN, (1981): Graphical data analysis, Annual Review of Psychology, Vol. 32, pp. 191-241.

UN PROGRAMME SAS POUR FAIRE DE L'ANALYSE FACTORIELLE DES CORRESPONDANCES

Bernard Clément, Ph.D,
Professeur titulaire
Département de Mathématiques appliquées
Ecole Polytechnique de Montréal

1- INTRODUCTION

Le système SAS (acronyme pour Statistical Analysis System) est un système informatique d'analyse statistique de données. Développé depuis une vingtaine d'années aux U.S.A., il est maintenant répandu mondialement et il est reconnu comme un des chefs de file des progiciels d'application. En plus de ses procédures statistiques d'excellente qualité, il s'est greffé au cours des récentes années, des prolongements en graphisme (SAS/GRAPH), en analyse économique (SAS/ETS), et en recherche opérationnelle (SAS/OR). C'est un des rares progiciels à pouvoir prétendre à l'intégration complète de toutes ses composantes. Développé originalement sur de gros ordinateurs IBM, on possède maintenant des versions sur des mini-ordinateurs VAX et des micro-ordinateurs PC, XT, AT.

Le système opère en mode lot (OS) ou en mode interactif sous CMS ou TSO. Il possède environ 150 procédures dont certaines telles PROC MATRIX, PROC GLM sont extrêmement puissantes de par la variété et la complexité de leurs options de commandes.

Malgré tout, le système SAS ne possédait pas de procédure pour effectuer l'analyse factorielle des correspondances. Cela s'explique en partie par le fait que l'analyse des correspondances est une technique d'analyse relativement peu connue aux U.S.A. Toutefois, cette lacune est en train de se corriger rapidement depuis quelques années par la parution de monographie en langue anglaise sur l'analyse de correspondances [3] [4].

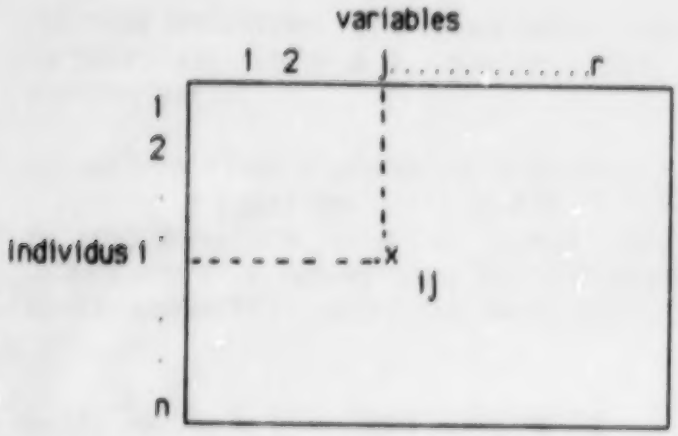
A l'été 1983, nous avons décider d'écrire un programme d'analyse de correspondances pouvant s'exécuter sous la supervision du système SAS. La première version de ce programme a été présentée au neuvième congrès annuel des utilisateurs du système SAS en 1984. Depuis notre programme a évolué et notre version actuelle permet de faire de l'analyse de correspondances multiples. Le mode d'utilisation du programme sera décrit plus loin et illustré à l'aide d'un exemple. Mais en premier lieu nous

présentons quelques éléments essentiels du système SAS et de l'analyse des correspondances.

2- LE SYSTEME SAS

Un concept fondamental du système SAS est la séparation entre l'étape de définition des données et l'étape d'analyse des données. En conséquence tout programme SAS est constitué d'une série de blocs appelés étape DATA et étape PROC. A l'étape DATA l'utilisateur définit ses données et le système constitue un fichier de type SAS; à l'étape PROC l'utilisateur invoque une procédure de la librairie des programmes SAS et identifie les options désirées et spécifiques à la procédure.

Le fichier SAS est l'interface entre les données brutes et la librairie des procédures. Le fichier SAS peut être temporaire, le temps de l'exécution d'un travail, ou il peut être permanent et stocké sur disque ou ruban magnétique. On peut représenter un fichier SAS comme un tableau rectangulaire croisant individus et variables. Les variables peuvent être de type numérique ou de type caractère.



Fichier SAS

A partir du moment où le fichier SAS a été correctement créé et devient une entité reconnue par le système, on peut appliquer n'importe laquelle des procédures de la librairie. Voici un petit exemple artificiel illustrant la structure d'un programme SAS suivi de quelques explications.

```
DATA NOTES ;
      INPUT        NOM $    PRENOM $ AGE
                   RES 1    RES 2    RES 3
      CARDS;
GOYER Bernard      12       80       60    71
BLOUIN Jean        12       64       71    83
DUPRE Anne         13       67       69    92
CLAVEAU Louis      11       39       48    60
MELOCHE Marc       13       49       79    30
;
PROC PRINT;
PROC MEANS;
PROC PLOT;
     PLOT  RES 1 * RES 2        RES 1*RES 3;
PROC CORR;
```

Explications

- Un fichier SAS portant le nom NOTES est à créer;
- les variables déclarées avec l'énoncé INPUT sont NOM PRENON AGE RES1 RES 2 RES 3;
- les variables NOM et PRENOM sont déclarées de type caractères avec le signe $.
- les autres variables sont de type numérique;
- l'énonce CARDS indique que les prochaines lignes contiennent les données;

- le fichier SAS contiendra 6 variables sur 5 individus;

- à la suite de cette première étape DATA, il vient 4 autres étapes PROC soient PRINT, MEANS, PLOT, CORR qui sont quelques unes des procédures disponibles;

- la fonction de ces 4 procédures est sommairement

  PRINT : impression des données d'un fichier SAS,

  MEANS : calcul des statistiques descriptives sur chacune des variables numériques du fichier,

  PLOT : trace des scattergrammes.

  CORR : calcule une matrice de corrélation.

la syntaxe minimale que l'utilisateur a besoin de connaître peut se résumer par quelques règles :

- les noms des fichiers SAS (étape DATA) ainsi que le nom des variables est limité à 8 caractères dont le premier est alphabétique;

- chaque énoncé SAS se termine par le symbole ";"

- le programme est écrit en format libre en respectant intégralement la nomenclature et l'orthographe des mots du vocabulaire de SAS comme DATA, PROC, INPUT, CARDS, etc...

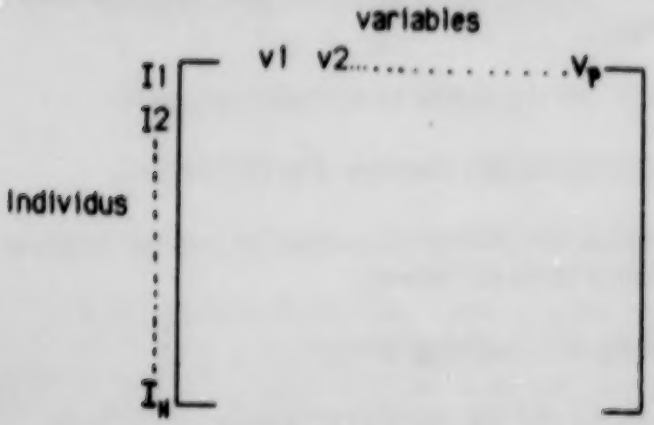
3- ANALYSE FACTORIELLE DES CORRESPONDANCES

On peut présenter l'analyse des correspondances de plusieurs points de vue : analyse en composantes principales selon la métrique du khi-deux, analyse canonique de deux paquets de variables logiques (indicatrices), codage optimal maximisant la corrélation entre deux variables qualitatives etc... De toutes les présentations il semble que celle décrivant l'analyse des correspondances comme une représentation simultanée des rangées et des colonnes d'un tableau de données dans des plans de variabilité maximale soit la meilleure.

Typiquement le tableau de données est un tableau de contingence croisant deux variables qualitatives. Toutefois la seule contrainte mathématique à

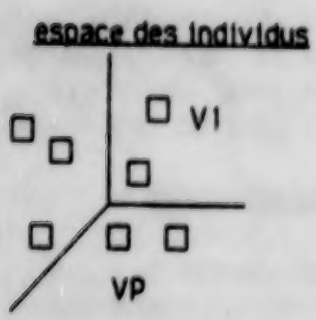
l'application de la méthode de factorisation est d'avoir un tableau de nombres positifs.

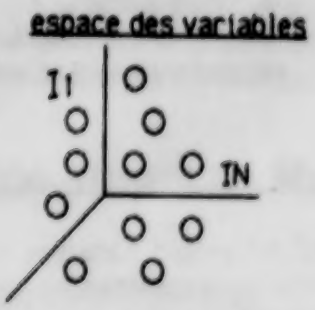
Tableau des données



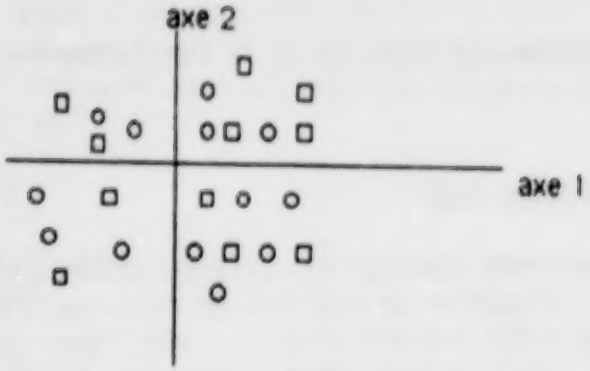
Le tableau peut être considéré de deux points de vue:

- P vecteurs-variables (colonnes) dans l'espace des individus en dimension N;
- N vecteurs-individus (rangées) dans l'espace des variables en dimension P;





De cette double représentation on obtient une représentation simultanée selon les principaux axes d'inertie



Le plan principal d'inertie permet une représentation des données dans un espace réduit de dimension 2 tout en conservant le maximum d'information sur la structure de celle-ci. Cela favorise la confection de scénarios pour l'interprétation et la perception des données.

La méthode d'analyse factorielle des correspondances se distingue de l'analyse factorielle classique en ce que la métrique employée est celle du khi-deux plutôt que la métrique euclidienne. D'autre part ce n'est pas le tableau des données brutes qui est soumis à l'analyse mais plutôt les profils c'est-à-dire les rangées (et les colonnes) normalisés par leurs sommes (distributions conditionnelles).

Sans entrer dans les détails de calculs, mentionnons que la recherche des axes d'inertie conduit à un problème matriciel de valeurs propres et vecteurs propres. Les vecteurs propres correspondent aux axes d'inertie auxquels sont associées des valeurs propres positives rangées dans l'ordre décroissant : $1 > \lambda_1 \geq \lambda_2 \geq \ldots\ldots \lambda_q > 0$. Dans un programme d'analyse de correspondance on présente les éléments suivants à la sortie :

- les valeurs propres, le pourcentage d'inertie de chaque axe ainsi que le pourcentage cumulatif;
- les coordonnées des points-lignes et des points-colonnes;
- les plans de projections avec points identifiés;
- la contribution des individus et la contribution des variables à chaque axe avec rangement des contributions dans l'ordre;

- la corrélation des individus et la corrélation des variables à chaque axe.

3- LE PROGRAMME SAS

Notre programme d'analyse des correspondances a été écrit en utilisant le langage de manipulation de matrices que constitue la procédure MATRIX de SAS. Le programme contient environ 1 000 lignes de code incluant les commentaires. Pour utiliser notre programme, l'utilisateur doit créer 7 fichiers SAS en amont du programme. Ces fichiers doivent obligatoirement porter les noms suivants : OBSERV, ICARAC, JCARAC, INAME, JNAME, OPTPR, OPTGR. Le rôle de chacun de ces fichiers est de définir la structure des données (OBSERV, ICARAC, JCARAC), les libellés des rangées et colonnes (INAME, JNAME) et de préciser les options de la sortie numérique ainsi que de la sortie graphique (OPTPR, OPTGR).

Voici la description sommaire de chacun de ces fichiers.

DATA ICARAC:
INPUT X ●● :
CARDS ;

IDIMP IDIMS
A1 A2 B1 B2 C1 C2.....

Signification des paramètres

IDIMP: Nombre de rangées principales
IDIMS: Nombre de rangées supplémentaires
(IDIMS=0 si aucune)

Les rangées supplémentaires (s'il y en a) sont identifiées par une suite de paires de numéros de rangées. Le programme contient un éditeur qui permute les rangées afin que les rangées principales viennent en premier suivies des rangées supplémentaires.

A1, A2: Les rangées du tableau de données (DATA OBSERV) entre la rangée A1 et la rangée A2 inclusivement sont supplémentaires.
B1,B2 : Même signification que A1, A2
C1, C2 : Même signification que B1,B2...

Par exemple si A1=8 A2=10 B1=13 B2=13 C1=41 C2=45
Les rangées 8, 9, 10 sont supplémentaires.

La rangée 13 est supplémentaire
Les rangées 41 42 43 44 45 sont supplémentaires et ne seront pas employées dans le calcul des valeurs propres et vecteurs propres.

DATA JCARAC :

INPUT X ●● :
CARDS :

JDIMP JDIMS
A1 A2 B1 B2 C1 C2.....

Signification des paramètres

JDIMP : Nombre de colonnes principales
JDIMS : Nombre de colonnes supplémentaires
(JDIMS=0 si aucune)

Les colonnes supplémentaires (s'il y en a) sont identifiées par une suite de paires de numéros de colonnes. Le programme contient un éditeur qui permute les colonnes afin que les colonnes principales viennent en premier suivies des colonnes supplémentaires.

A1, A2 :Les colonnes du tableau de données (DATA OBSERV) entre la colonne A1 et la colonne A2 inclusivement sont supplémentaires.
B1, B2 : Même signification que A1, A2
C1,C2 : Même signification que B1, B2...

Par exemple si A1=8 A2=10 B1=13 B2=13 C1=41 C2=45.
Les colonnes 8 9 10 sont supplémentaires.
La colonne 13 est supplémentaire.
Les colonnes 41 42 43 44 45 sont supplémentaires et ne seront pas employées dans le calcul des valeurs propres et vecteurs propres.

DATA OBSERV

Le tableau des données.
Le tableau doit satisfaire les deux conditions suivantes :

- Nombres réels positifs ou nuls
- Au moins un nombre positif dans chaque rangée et colonne.

DATA INAME

Les libellés des rangées. On recommande d'employer 3 caractères par libellé. Les libellés ayant plus de 3 caractères seront tronqués. Ils doivent être dans l'ordre des rangées de data OBSERV.

DATA JNAME

Les libellés des colonnes. On recommande d'employer 3 caractères par libellé. Les libellés ayant plus de 3 caractères seront tronqués. Ils doient être dans l'ordre des colonnes de data OBSERV.

DATA OPTPR :
INPUT X ●● :
CARDS :

SP LNINCH P1 P2 P3.....

<u>Paramètres de la sortie numerique</u>

SP=1 impression du tableau des données (OBSERV)
= 0 pas d'impression du tableau des données

LNINCH = nombre de lignes au pouce sur l'imprimante (habituellement 8 ou 10)

P1=1 par axe. Les coordonnées négatives viennent en premier suivies des coordonnées positives. Les vecteurs sont ordonnés selon l'ordre décroissant des CTREL à l'intérieur de chaque groupe. Les CTREL sont employés pour définir les axes.
=2 coordonnées simultanées sur les six premiers axes.
=3 option 1 et 2 précédentes combinées.

P2 = blanc (avec P1=2)
= nombre de l'axe désiré (avec P1=1 ou P1=3)
= 0 tous les axes

P3, P4... même signification que P2

DATA OPTGR :
INPUT X1-X7 :
CARDS :

AXEH AXEV TYPEI TYPEJ GPAGE GLINE SWFAR

Paramètres de la sortie graphique

AXEH : Numéro du facteur sur l'axe horizontal

AXEV : Numéro du facteur sur l'axe vertical

TYPEI=1 rangées principales seulement
=2 rangées supplémentaires seulement
=3 toutes les rangées
=0 autrement

TYPEJ=1 ou 2 ou 3 ou 0
même signification que TYPEI mais pour les colonnes

GPAGE =1 Graphique sur une page de largeur
=2 graphique sur deux pages de largeur

GLINE = 0 Graphique avec même échelle sur les deux axes.
= N 30<=N<=250 Nombre de lignes horizontales
SWFAR= 1 Les points à plus de trois écarts-types sont placés sur le cadre
= 0 autrement

Un graphique est spécifié à l'aide de ces 7 paramètres. Plusieurs graphiques peuvent être obtenus en spécifiant autant de lignes de paramètres que l'on désire.

4- UN EXEMPLE

Les données étaient constituées d'un tableau de contingence de dimension 27 x 12 croisant 27 mots employés dans le vocabulaire de 12 candidats à une élection à Rennes. Le tableau de données avait été utilisé par Mme Brigitte Escofier lors de sa présentation sur l'analyse des correspondances.

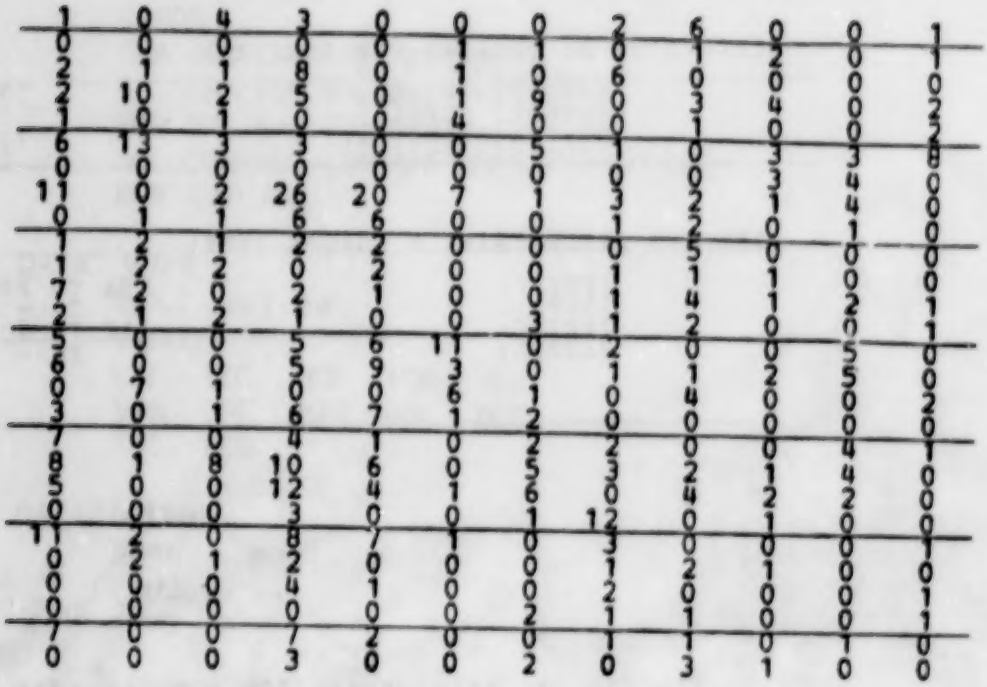
Tableau des données

LAP	LEC	HER	POR	GUI	VAN	CRE	HAR	ROL	DEC	TEC	GOR	
CHA	1.	0.	4.	3.	0.	0.	0.	2.	6.	0.	0.	1.
PRO	0.	0.	0.	0.	0.	1.	1.	0.	1.	2.	0.	1.
DRO	2.	1.	1.	8.	0.	1.	0.	6.	0.	0.	0.	0.
LIB	2.	10.	2.	5.	0.	1.	9.	0.	3.	4.	0.	2.
IND	1.	0.	1.	0.	0.	4.	0.	0.	1.	0.	0.	2.
DEM	6.	13.	3.	3.	0.	0.	5.	1.	0.	3.	3.	8.
DIC	0.	0.	0.	0.	0.	0.	0.	0.	0.	3.	4.	0.
TRA	11.	0.	2.	26.	20	7.	1.	3.	2.	1.	4.	0.
FEM	0.	1.	1.	6.	6.	0.	0.	1.	2.	0.	1	0.
AL	1.	2.	1.	2.	2.	0.	0.	0.	5.	0.	0.	0.
PRI	1.	2.	2.	0.	2.	0.	0.	1.	1.	1.	0.	0
CHO	7.	2.	0.	2.	1.	0.	0.	1.	4.	1.	2.	1.
EMP	2.	1.	2.	1.	0.	0.	3.	1.	2.	0.	0.	1.
DIT	5.	0.	0.	5.	6.	11.	0.	2.	0.	0.	5.	0.
GAU	6.	0.	0.	5.	9.	3.	0.	1.	1.	2.	5.	0.
FAM	0.	7.	1.	0.	0.	6.	1.	0.	4.	0.	0.	2.
LON	3.	0.	1.	6.	7.	1.	2.	0.	0.	0.	0.	0.
CRI	7.	0.	0.	4.	1.	0.	2.	2.	0.	0.	4.	1.
SOC	8.	1.	8.	10.	6.	0.	5.	3.	2.	1.	4.	0.
COM	5.	0.	0.	12.	4.	1.	6.	0.	4.	2.	2.	0.
AUT	0.	0.	0.	3.	0.	0.	1.	12.	0.	1.	0.	0
LUT	11.	2.	0.	8.	7.	1.	0.	3.	0.	0.	0.	1.
SMI	0.	2.	1.	2.	0.	0.	0.	1.	2.	1.	0.	0.
NUC	0.	0.	0.	4.	1.	0.	0.	2.	0.	0.	0.	1.
REG	0.	0.	0.	0.	0.	0.	2.	1.	1.	0.	0.	1.
CAP	7.	0.	0.	7.	2.	0.	0.	1.	1.	0.	1.	0.
PRL	0.	0.	0.	3.	0.	0.	2.	0.	3.	1.	0.	0.

Les colonnes sont libellées avec les trois premières lettres du nom des candidats : LAF, LEC, ... GOR tandis que les rangées sont les mots employés CHA (changement), PRO (programme),..., PRC.

Pour effectuer l'analyse des correspondances on soumet pour exécution les énoncés SAS suivants suivis du programme SAS d'analyse de correspondances.

```
DATA OBSERV;
      INPUT N●●;
      CARDS;
```



```
DATA ICARAC;
INPUT I ●● ;
CARDS ;
27 0
```

DIMENSIONS DU TABLEAU DES DONNEES:

RANGEES PRINC.:	27
RANGEES SUPPL.:	0
COLONNES PRINC.:	12
COLONNES SUPPL.:	0

LIBELLES APPARTENANT A CHAQUE TYPE:

IPRINC:	CHA TO PRC
ISUPPL:	- TO -
JPRINC:	LAF TO GCH
JSUPPL:	- TO -

TABLEAU DES VALEURS PROPRES ACTIVES
(LA VALEUR PROPRE "1" EST ELIMINEE)

NO	LAMBDA	POURCENT	CUMUL
1	0.36419	25.808	25.808
2	0.25389	17.991	43.799
3	0.17990	12.741	56.540
4	0.17355	12.298	68.838
5	0.12464	8.832	77.671
6	0.09667	6.141	83.812
7	0.07521	5.329	89.142
8	0.06085	4.312	93.454
9	0.03888	2.755	96.209
10	0.02885	2.044	98.253
11	0.02465	1.747	100.000
	1.41118	100.000	

```
DATA JCARAC ;
     INPUT J ●● ;
     CARDS ;
     12 0
     ;
DATA INAME ;
     INPUT IN$ ●● ;
     CARDS ;
     CHA PRO DRO LIB IND DEM
     DIC TRA FEM SAL PRI CHO
     EMP OIT GAU FAM CON CRI
     SOC COM AUT LUT SMI NUC
     REG CAP PRC
     ;
DATA JNAME ;
     INPUT JN$ ●● ;
     CARDS ;
     LAF  LEC  HER  POR  GUI
     VAN  CRE  MAR  ROL  DEC
     TEC  GOR
     ;
DATA OPTPR ;
     INPUT A ●● ;
     CARDS ;
     1 8 3 1 2
     ;
DATA OPTGR ;
     INPUT  AXEH  AXEV  TYPEI  TYPEJ
            GPAGE GLINE SWFAR ;
     1  2  3  3  1  0  1
     ;
PROC MATPRIX ;
```

Programme SAS d'environ 1000 lignes

Voici ce que donne l'exécution du programme à la sortie :

COORDONNEES SELON L'ORDRE DECROISSANT DE CTREL
(NEGATIVES ET POSITIVES)

(INER % COORD 100:1 CTREL % COS2 100:1)

AXIS=1 VARTYP=IPRINC

VAR	MASS	INER	COORD	CTREL	COS2
DEM	0.[illegible]4	8.6	-101.9	21.2	63.3
LIB	0.063	6.3	-102.4	18.1	73.7
FAM	0.035	6.6	-106.2	10.7	42.1
PRO	0.010	2.8	-97.2	2.6	24.0
EMP	0.021	1.7	-65.3	2.2	19.0
REG	0.008	2.1	-89.8	1.3	32.3
SMI	0.015	1.4	-63.7	1.7	30.7
CHA	0.028	4.5	-38.6	1.1	6.6
SAL	0.021	2.7	-42.1	1.0	9.3
PRC	0.015	2.2	-47.1	0.9	10.6
IND	0.015	4.3	-43.6	0.8	4.6
PRI	0.017	1.5	-41.0	0.8	13.1
CHO	0.035	1.9	-15.1	0.2	3.0
DIC	0.012	7.8	-14.1	0.1	0.2
TRA	0.127	4.3	59.3	12.3	73.3
OIT	0.056	6.5	57.2	5.0	19.9
GAU	0.053	3.7	54.8	4.4	42.6
LUT	0.055	2.5	45.2	3.1	32.0
CAP	0.031	1.8	55.7	2.7	37.6
CON	0.033	2.0	52.6	2.5	32.9
AUT	0.028	12.2	68.5	1.8	3.8
DRO	0.031	3.2	41.3	1.5	12.2
FEM	0.030	1.7	40.7	1.4	20.4
CRI	0.035	2.3	30.9	0.9	10.3
NUC	0.013	1.5	41.4	0.6	10.5
SOC	0.079	2.7	10.0	0.2	2.1
COM	0.060	2.1	9.1	0.1	1.6

COORDONNEES SELON L'ORDRE DECROISSANT DE CTREL
(NEGATIVES ET POSITIVES)

(INER % COORD 100:1 CTREL % COS2 100:1)

AXIS=1 VARTYP=JPRINC

VAR	MASS	INER	COORD	CTREL	COS2
LEC	0.073	11.6	-124.2	30.8	68.6
GOR	0.036	6.8	-113.5	12.9	48.3
CRE	0.066	7.4	-71.5	9.3	32.3
ROL	0.074	8.9	-56.6	6.6	18.9
DEC	0.038	7.4	-52.9	4.1	14.4
HER	0.050	5.6	-44.1	2.7	12.3
GUI	0.122	7.7	68.3	15.7	52.7
POR	0.207	5.1	39.9	9.0	46.0
MAR	0.073	15.3	42.7	3.6	6.1
LAF	0.142	5.2	29.3	3.3	16.5
TEC	0.058	7.5	32.3	1.7	5.7
VAN	0.061	11.6	15.5	0.4	0.9

COORDONNEES SELON L'ORDRE DECROISSANT DE CTREL
(NEGATIVES ET POSITIVES)
(INER % COORD 100:1 CTREL % COS2 100:1)

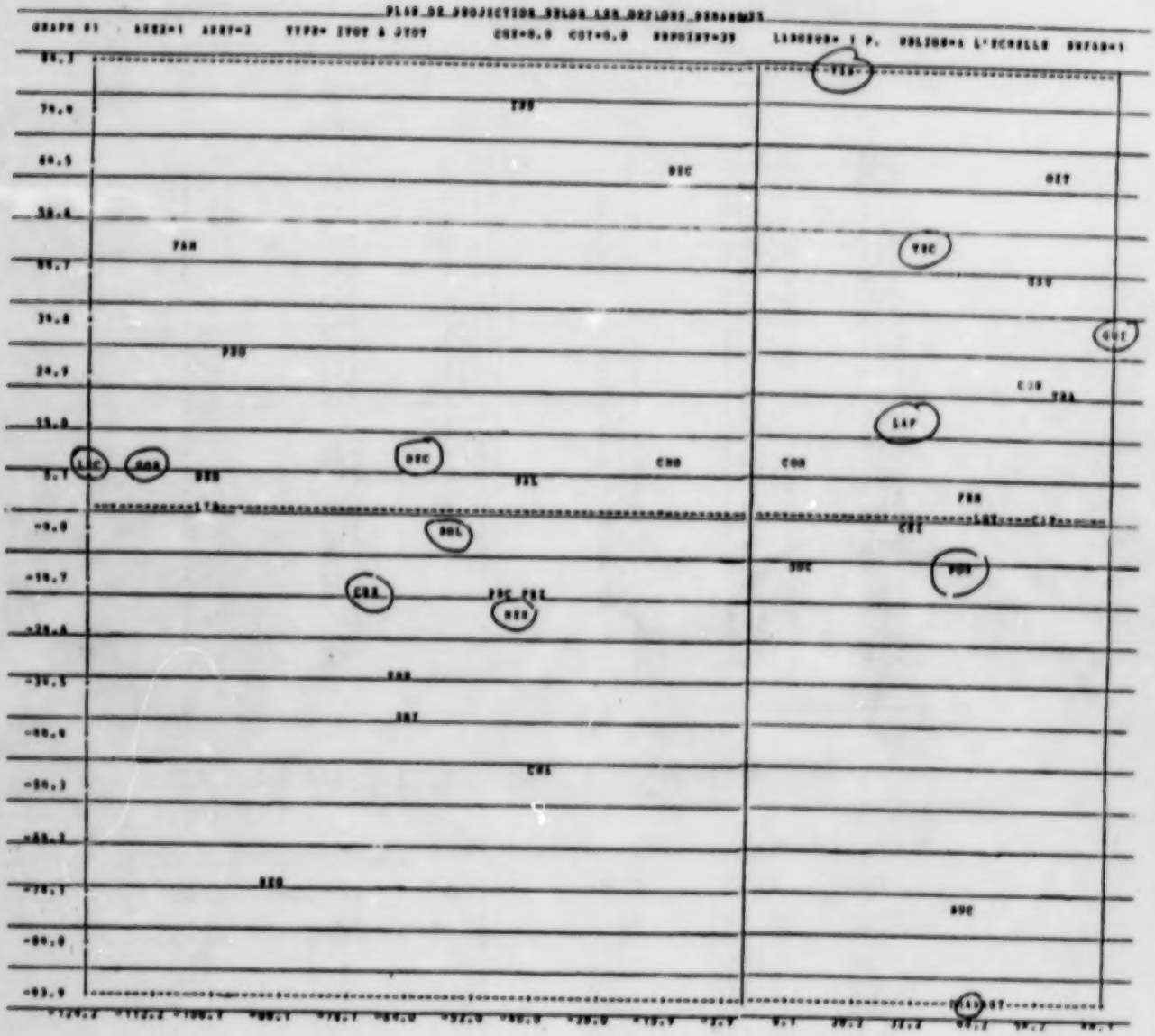
----------AXIS=2 VARTYP=IPRINC----------

VAR	MASS	INER	COORD	CTREL	COS2
AUT	0.028	12.2	-218.3	52.8	77.6
DRO	0.031	3.2	-93.9	10.9	[illegible]
NUC	0.013	1.5	-75.9	3.0	[illegible]
CHA	0.028	4.5	-49.9	2.8	[illegible]
REG	0.008	2.1	-72.6	1.7	1.1
SNI	0.015	1.4	-40.2	0.9	14.6
EMP	0.021	1.7	-33.1	0.9	12.2
SOC	0.079	2.7	-10.8	0.4	2.8
PRI	0.017	1.5	-17.6	0.2	2.5
PRC	0.015	5.2	-16.1	0.2	2.4
CRI	0.035	2.3	-2.0	0.0	1.2
LIB	0.063	6.3	-1.1	0.0	0.0
LUT	0.055	2.5	-0.9	0.0	0.0
CAP	0.011	1.8	-0.0	0.0	0.0
OIT	0.056	6.5	65.1	9.4	25.2
GAU	0.053	2.7	45.1	4.2	28.7
IND	0.015	4.3	75.6	3.3	14.0
FAM	0.035	6.6	48.5	3.2	8.8
TRA	0.127	4.3	22.5	2.5	10.6
DIC	0.012	7.8	65.4	1.9	4.5
CON	0.033	2.0	25.8	0.9	7.9
PRO	0.010	2.8	28.9	0.3	2.1
COM	0.060	2.1	9.6	0.2	1.8
CHO	0.035	1.9	8.1	0.1	0.9
DEM	0.074	8.6	4.6	0.1	0.1
SAL	0.021	2.7	4.9	0.0	0.1
FEM	0.030	1.7	2.2	0.0	0.1

COORDONNEES SELON L'ORDRE DECROISSANT DE CTREL
(NEGATIVES ET POSITIVES)
(INER % COORD 100:1 CTREL % COS2 100:1)

----------AXIS=2 VARTYP=JPRINC----------

VAR	MASS	INER	COORD	CTREL	COS2
MAR	0.073	15.3	-152.5	66.6	79.4
POR	0.207	5.1	-11.0	1.0	3.5
HER	0.050	5.6	-20.4	0.8	2.6
CRP	0.066	7.4	-16.4	0.7	1.7
ROL	0.074	8.9	-5.4	0.1	0.2
VAN	0.061	11.6	84.3	17.1	26.5
GUI	0.122	7.7	35.2	6.0	14.0
TEC	0.058	7.5	51.2	6.0	14.4
LAF	0.142	5.2	16.1	1.5	5.0
LEC	0.073	11.6	7.3	0.2	0.5
DEC	0.038	7.4	8.6	0.1	0.3
GOR	0.036	6.8	7.7	0.1	0.2


PLAN DE PROJECTION SELON LES OPTIONS DEMANDEES
GRAPH 01 : AXEX=1 AXEY=2 TYPE= IPOT A JPOT CXE=0.0 CXT=0.0 DEPOINT=39 LANGUE= 1 P. POLICE=1 L'ECHELLE DEFAU=1
89.3
79.4
69.5
59.6
49.7
39.8
29.9
20.0
10.1
0.0
-9.8
-19.7
-29.6
-39.5
-49.4
-59.3
-69.2
-79.1
-89.0
-93.9
-124.2
-112.2
-100.1
-88.1
-76.1
-64.0
-52.0
-40.0
-28.0
-16.0
-4.0
8.1
20.1
32.2
44.2
56.2
68.2
80.2
92.3

BIBLIOGRAPHIE

1. Bernard Clément (1984), Introduction à SAS (Statistical Analysis System), Edition de l'école polytechnique de Montréal, 108 pages.

2. Bernard Clément, Daniel Lévesque (1984), A SAS Program for Correspondance Analysis Using PROC MATRIX Proceeding of the inth Annual SVGI. Conference (SAS Users Group International) Florida, March 18-21, 1984. pages 645-650.

3. Micheal J. Greenacre (1984), Theory and Applications of Correspondance Analysis. Academic Press.

4. Lebart L., Morineau A., and Warlvick K.M. (1984), Multivariate descriptive Statistical Analysis, John Wiley Inc., New York.

Ateliers d'analyse de données didactiques

Les activités présentées dans les deux ateliers avaient le même objectif : donner l'occasion d'appliquer les méthodes présentées dans les exposés théoriques, ceci en vue de favoriser non seulement une meilleure appropriation mais également une démythification de la branche des mathématiques appliquées : "analyse des données".

ATELIER 1: EXEMPLES DE TRAITEMENT DE DONNEES PAR ANALYSE DES CORRESPONDANCES

Personne-ressource : Brigitte Escofier

L'atelier sur la méthode d'analyse des correspondances a été organisé à partir de deux exemples de données, l'un fourni par deux chercheures de l'UQAM, l'autre par le Service de consultation en analyse de données (SCAD). Les participants de l'atelier ont été amenés à parcourir, à partir de ceux-ci, les différentes étapes d'une analyse et ainsi à les saisir davantage.

Nous reprenons les principales questions soulevées lors de cet atelier au fur et à mesure du travail sur ces deux exemples, et présentons partiellement les deux exemples analysés.

I / EXEMPLE 1

Il s'agit d'une analyse des correspondances simple. Les données sont tirées d'une recherche sur le langage effectuée par Helga Felder (Département de psychologie, UQAM) et Sylvie Cantin (étudiante au doctorat en psychologie, UQAM).

Voici l'hypothèse de la recherche à partir de laquelle les données qui suivent (voir tableau 1) ont été recueillies. Partant d'une théorie des actes du langage permettant d'établir une taxonomie ou un système de classification des verbes communicatifs (sens interpersonnels des verbes), on cherche à tester empiriquement si ces catégories théoriques correspondent à la perception des gens.

Les 20 verbes présentés correspondent selon la théorie sous-jacente à quatre catégories interlocutoires (quatre catégories d'actions interpersonnelles). La perception des gens confirme-t-elle cette catégorisation?

1. Collecte des données

Les 20 verbes de la liste initiale ont été groupés par paires, par exemple: affirmer, commander, affirmer, confirmer,...etc.

380 paires ont ainsi été formées et présentées dans un questionnaire. Les répondeurs devaient évaluer, à l'aide d'une échelle de 1 à 9, le degré de similarité entre les deux verbes présentés.

Le tableau suivant a été constitué à partir des réponses au questionnaire.

	AFF	COM	CON	DEM	FEL	GAR	INF	INT	JUG	OFF	ORD	PAR	PRO	REC	REG	REM	SEN	SEX	SOU	SUP
AFF	81	44	60	41	40	52	61	33	49	33	42	56	47	47	34	34	54	21	34	42
COM	46	81	42	37	20	28	37	42	30	21	46	19	26	62	18	22	20	16	28	23
CON	60	42	81	39	39	62	68	33	40	40	23	38	49	27	22	37	50	27	33	45
DEM	41	37	39	81	25	26	41	37	29	42	41	25	29	46	29	41	28	34	41	34
FEL	40	20	39	25	81	35	19	13	41	38	17	29	30	18	35	33	23	37	56	30
GAR	52	28	62	26	35	81	46	28	39	48	25	49	69	34	21	31	74	38	40	28
INF	61	37	68	41	19	46	81	30	37	37	30	40	36	27	27	24	50	28	33	37
INT	33	42	33	37	13	28	30	81	38	24	47	20	27	50	24	20	26	16	17	23
JUG	49	30	40	29	41	39	37	38	81	26	47	39	28	32	32	37	38	37	31	41
OFF	33	21	40	42	38	48	37	24	26	81	27	43	47	20	33	46	32	40	36	36
ORD	47	46	23	41	17	25	30	47	47	23	81	25	15	44	10	19	24	17	20	19
PAR	56	19	38	25	29	49	40	20	39	43	25	81	49	24	33	26	39	29	41	63
PRO	47	26	49	29	30	69	36	27	28	43	15	49	81	15	43	36	72	44	39	43
REC	47	46	27	46	18	34	27	50	32	20	44	24	15	81	21	17	27	21	28	17
REG	34	18	33	29	35	21	27	24	32	35	10	33	43	21	81	39	33	72	43	28
REM	34	22	37	41	33	31	24	20	37	46	19	26	36	17	39	81	38	56	52	23
SEN	54	20	50	28	23	74	50	26	38	32	24	39	72	27	33	38	81	33	30	27
SEX	21	16	27	34	37	38	28	16	37	40	17	29	44	21	72	56	35	81	42	30
SOU	34	28	35	41	56	40	33	17	31	56	20	41	39	28	43	32	30	42	81	32
SUP	42	23	45	34	30	28	37	23	41	36	19	63	43	17	28	23	27	30	32	81

AFF : AFFIRMER
COM : COMMANDER
CON : CONFIRMER
DEM : DEMANDER
FEL : FELICITER
GAR : GARANTIR
INF : INFORMER
INT : INTERDIRE
JUG : JUGER
OFF : OFFRIR
ORD : ORDONNER
PAR : PARIER
PRO : PROMETTRE
REC : RECLAMER
REG : REGRETTER
REM : REMERCIER
SEN : S'ENGAGER
SEX : S'EXCUSER
SOU : SOUHAITER
SUP : SUPPOSER

Tableau 1

Quelques remarques sur les données

Il ne s'agit pas ici d'un tableau habituel de données (nous n'avons pas affaire, en effet, à un tableau de comptage, de contingence). De plus, l'analyse porte ici directement sur la ligne et non sur le profil (déterminé en prenant les nombres divisés par la somme de toute la ligne).

Dans ce cas-ci, toutefois, définir l'indice de proximité sur la ligne ou le profil ne fait pas une grande différence. Dans le cas de la ligne, les données concernées sont les chiffres bruts. Ainsi, les deux verbes sont proches si les lignes sont proches, i.e. si les chiffres bruts se ressemblent.

Question 1: Quels sont les avantages de l'analyse des correspondances dans ce cas, par rapport à d'autres méthodes?

Réponse: L'analyse des correspondances effectuée à partir des résultats du questionnaire vient confirmer la théorie. Elle permet de faire ressortir les regroupements les plus significatifs, et permet une visualisation graphique de ces catégories.

Par rapport à la méthode "Multiple Scales" également utilisée précédemment dans l'analyse, elle nous donne une photo simultanée en 2 dimensions, laquelle met en évidence les composantes principales, alors qu'avec l'analyse "Multiple Scales", on pouvait obtenir des résultats en 5 ou 6 dimensions, ce qui posait évidemment des problèmes d'interprétation.

Question 2: A la question "Si on vous avait présenté les données brutes, quelle aurait été votre démarche?"

Réponse: Mme Escofier confirme qu'elle aurait fait une analyse des correspondances, parce que dans des cas similaires cette analyse s'est montrée très efficace. Toutefois, les sommes des lignes étant ici très proches, les profils la gênent quelque peu.

Sur de petits tableaux (ce qui est le cas ici), Mme Escofier ne fait quasiment que de l'analyse des correspondances. Par contre, sur de grands tableaux, (par exemple une enquête auprès de 500,1000 individus...), elle fait simultanément plusieurs choses:

analyse des correspondances et classification. Les données étant alors beaucoup plus complexes, il est intéressant de les envisager sous plusieurs aspects. Lorsqu'on a énormément d'individus, les techniques de classification et de représentations synthétiques sont très efficaces. Par contre, une analyse de classification est rarement intéressante sur un petit nombre de données.

Question 3: La matrice de données est ici une matrice symétrique. Cela influe-t-il sur l'analyse des correspondances?"

Réponse: Il y a forcément des résultats spécifiques aux matrices symétriques, puisque les lignes et les colonnes ont nettement la même structure, souligne Mme Escofier.

Les valeurs qu'on met sur la diagonale (ici 81, un verbe ayant reçu face à lui-même la côte de similarité maximum : 9) influent assez peu sur les facteurs contribuant à la formation des axes, autrement dit, les facteurs restent proportionnels. En effet que l'on mette 0 ou 81, il n'y a que les valeurs propres qui changent.

Question 4: Quels autres types de tableaux de données (en dehors des tableaux habituels de comptage, de contingence) peut-on traiter par analyse des correspondances?

Réponse: Des tableaux de note, des tableaux de mesures numériques peuvent être traités par analyse des correspondances en les dédoublant.

Par exemple, pour un tableau de notes situées entre 0 et 20, on dédouble chaque note en faisant une colonne pour la note n et une colonne pour 20 - n.

Des tableaux de codages disjonctifs complets peuvent être traités et, très fréquemment aussi, des tableaux de fréquences juxtaposées. Si on s'intéresse par exemple aux liens entre une variable et trois autres variables, il faut construire un tableau qui met en ligne les modalités de la variable qui doit être expliquée et en colonne les modalités de chacune des autres

variables. Il ne s'agit pas alors exactement d'un tableau de fréquences.

Il n'y a pas de limite dans le nombre de variables, mais par contre, dans les modalités de la variable on ne conseille pas plus que 4,5 modalités de variable.

Question 5: Relativement à cette juxtaposition de tableaux de fréquences très utilisée en analyse des correspondances, comment interprétez-vous les axes dans ce cas-ci? Qu'est-ce que vous regardez...? Procédez-vous de la même façon que dans le cas d'une analyse de correspondances classique, ou n'y a-t-il pas quelque variante dans la façon de regarder les résultats?

Réponse: Globalement c'est la même chose, bien qu'on ait en tête des références supplémentaires qui font que ce sont des tableaux de juxtaposition.

L'interprétation des proximités est faite de la même façon... Tout comme dans la démarche normale, on regarde où se situe un point x par rapport à l'ensemble des données.

2. Interprétation des résultats (de l'exemple 1)

Question 1: La difficulté ne se situe pas, dans l'analyse des correspondances, au niveau du calcul; les programmes sont là pour les faire, elle se situe plus au niveau de l'interprétation. Ce qui manque le plus à cet effet, ce sont des guides pour aider à regarder les indicateurs intéressants; les systèmes experts sont peut-être une solution future dans ce sens. A part Fenelon qui a écrit un livre sur ce sujet, y a t-il d'autres exposés sur cette question?

Réponse: Mme Escofier confirme effectivement le fait qu'il y a peu d'articles ou de références sur l'interprétation de l'analyse des données et qu'il faut comprendre à partir d'exemples.

Question 2: N'y aurait-il pas toutefois une marche à suivre au moins

minimale qui pourrait être précisée, une sorte de guide reprenant les différentes étapes à suivre?

Réponse: La complexité de l'écriture d'un tel guide est grande. La connaissance du type de données à analyser est requise pour écrire une marche à suivre. Il ne peut y avoir de marche à suivre sans considération à la fois de la problématique du chercheur et de la structure de ses données au sens large du terme. Il est, autrement dit, très difficile de formaliser tous les types de données et de problématiques qui peuvent se présenter.

Question 3: N'y a t-il pas une distinction importante à faire entre deux types de données reliées aux problématiques actuelles des recherches sur l'apprentissage ?

En effet, deux catégories importantes de données sont souvent envisagées: celles où des hypothèses sont posées à priori, cas des données présentées dans le premier exemple (système de classification théorique des verbes qu'on veut confirmer), celles, au contraire, de type descriptif dans lesquelles des facteurs principaux ou des grandes composantes sont à dégager et, à partir desquelles éventuellement des hypothèses seront formulées.

Réponse: Mme Escofier considère fausse en pratique cette distinction entre les deux catégories de recherches. A partir du moment où un tableau de données est construit, il est sûr que des idées sur celles-ci sont déjà présentes pour orienter.

Lorsque deux axes différents sont présents, cela permet de décrire une dispersion qui est due à tel ou tel effet... Dans quel cadre, une telle analyse entre-t-elle? Est-ce descriptif? La distinction entre les deux catégories est en pratique artificielle. L'analyse des correspondances peut servir à confirmer une hypothèse très précise comme dans ce premier exemple. Souvent, elle peut servir non seulement à vérifier des hypothèses, mais aussi à comparer un certain nombre d'hypothèses et à mesurer l'importance de l'une par rapport à l'autre.
Par contre, parfois il est possible de se servir de l'analyse des

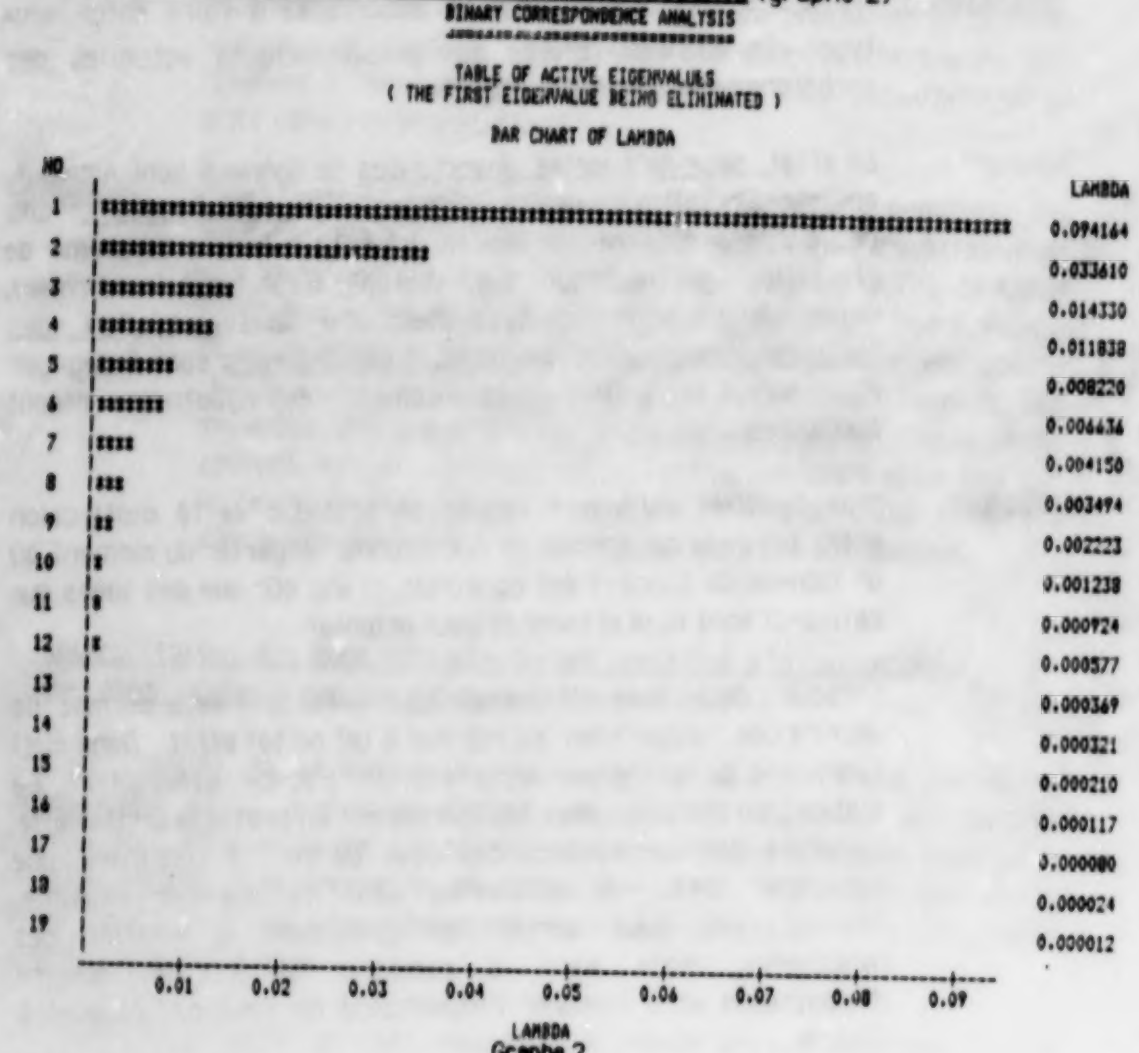
correspondances pour extraire un certain nombre de variables plus pertinentes (c'est également dans ce cas une question très précise avec une réponse précise).

En ce qui concerne l'exemple présenté ici, l'idée était de confirmer un modèle théorique très précis.

3. Différentes étapes de l'interprétation des résultats

Nous reprenons ici l'étude des résultats de l'exemple 1.

- Les valeurs propres : nombre d'axes intéressants (graphe 2)



Graphe 2

Question 1: Où s'arrête-t-on dans le nombre d'axes?

Réponse: Il semble qu'il y ait ici (graphe 2) deux facteurs importants. L'un des deux est vraiment prépondérant, après il y a un décalage et une baisse.

En fait, en général, il n'y a pas de réponse systématique à la question "où s'arrête-on"? En pratique quand l'interprétation des facteurs devient impossible l'analyse s'interrompt.

Question 2: A la lecture des livres de Benzécri, celui-ci semble aller jusqu'au 4e, 5e et 6e axe dans son interprétation, et ce avec beaucoup de facilité, alors que beaucoup de chercheurs s'en tiennent à l'interprétation des 2 premiers axes. Jusqu'où peut-on aller dans l'interprétation?

Réponse: Mme Escofier fait remarquer que cela dépend beaucoup des données (quantité de données) d'une part et de l'investissement mis dans les données. Si c'est une enquête qui a demandé énormément de temps, qui est extrêmement riche, il peut y avoir interprétation de beaucoup d'axes, mais dans un tableau comme celui-ci, un ou deux facteurs, c'est déjà suffisant. Pour 10 000 individus, 200 questions, cela peut être différent. Il nous arrive souvent, en fait, d'interpréter un facteur, un autre et de voir que le 3e est inintéressant. En bref, il semble donc que le nombre d'axes intéressants soit relatif à la richesse des données.

- <u>Interprétation des axes, facteurs qui ont contribué à la formation des axes : colonne des contributions.</u> (voir tableau 3)

Le logiciel utilisé dans cette analyse prévoit une sortie par axe où les coordonnées sont ordonnées (allant des parties négatives aux parties positives). Cette sortie par axe précise tout de suite les éléments importants qui contribuent à la formation de l'axe. Il prévoit également une sortie simultanée, celle que nous présentons (tableau 3) qui est la sortie simultanée des 5 - 6 premiers axes.

Pour interpréter les axes, nous faisons l'étude des contributions. Sur le graphe (tableau 3), par exemple l'étude de la colonne CTREL1 fait apparaître 4

éléments prédominants (ils forment à eux quatre au total à peu près 70% de l'axe) : commander, interdire, ordonner, réclamer. Ce sont ces 4 éléments qui ont formé l'axe, ce qui constitue la base de l'interprétation.

L'examen des coordonnées (COORD 1), montre que les quatre verbes ont des coordonnées négatives.

Pour le 2e axe (CTREL 2), il y a 3 facteurs qui ressortent, formant à eux trois 32% de l'inertie de l'axe : regretter, remercier, s'excuser qui contribuent respectivement pour 10.7%, 10.6% et 12% de l'inertie de l'axe pourcentage important puisqu'on a ici 20 verbes (ce qui correspond à une moyenne de 5%, si ceux-ci avaient tous la même contribution à l'inertie). Ce sont des verbes expressifs de reconnaissance, de politesse.

Le pourcentage total, 32%, est cependant ici moins fort que dans le cas du 1er axe et donc, c'est aussi le support visuel du graphique qui est important dans l'analyse de ce 2e axe (voir graphe 4).

BINARY CORRESPONDENCE ANALYSIS

EDITING OF COORDINATES

(INER % COORD 100:1 CTREL % COS2 100:1)

VARTYP=IPRINC

VAR	MASS	INER	COORD 1	COORD 2	COORD 3	COORD 4	COORD 5	COORD 6	CTREL 1	CTREL 2	CTREL 3	CTREL 4	CTREL 5	CTREL 6	COS2 1	COS2 2	COS2 3	COS2 4	COS2 5	COS2 6
AFF	0.060	1.7	-7.6	15.1	-5.6	2.4	3.9	5.6	0.4	4.1	1.3	0.3	1.1	2.9	11.4	44.3	6.0	1.1	3.0	6.2
CON	0.047	9.1	-37.5	-3.9	2.3	2.0	-0.7	-0.6	16.2	0.3	0.2	0.2	0.0	0.0	94.1	1.0	0.2	0.1	0.0	0.0
CON	0.057	2.3	3.0	19.5	-0.8	9.0	-1.9	13.0	0.1	6.4	0.0	3.9	0.2	14.6	1.3	52.6	0.1	11.1	0.5	23.5
DEM	0.049	3.0	-12.7	-12.7	-2.0	4.9	-23.9	4.1	0.8	2.3	0.1	1.0	34.2	1.2	14.3	14.2	0.3	2.1	30.6	1.5
FEL	0.045	5.6	27.0	-22.9	-14.7	21.8	16.3	0.8	3.5	6.9	6.7	17.9	14.4	0.0	31.8	22.7	9.4	20.6	11.5	0.0
GAR	0.056	3.2	12.1	22.5	14.6	6.4	7.4	-6.0	0.9	8.5	9.3	1.9	3.7	3.0	14.0	48.3	20.3	3.9	5.3	3.4
INF	0.052	3.0	-1.3	22.9	-0.4	1.7	-10.3	18.8	0.0	8.1	0.0	0.1	6.7	27.4	0.2	49.2	0.3	9.9	33.0	
INT	0.045	9.3	-38.2	-6.3	5.6	-7.6	1.7	-1.8	14.1	0.6	1.0	2.2	0.2	0.2	89.4	1.1	0.8	1.8	0.1	0.1
JUG	0.051	2.4	-6.5	-2.9	-11.8	-4.4	19.0	7.7	0.5	0.1	4.9	0.8	22.3	4.6	7.0	1.0	16.4	2.3	42.4	7.0
OFF	0.052	3.0	22.1	-2.3	4.1	9.0	-12.7	-12.0	2.7	0.1	0.6	3.3	10.0	11.2	45.5	0.3	1.6	7.3	14.9	11.5
ORD	0.046	11.5	-66.1	-8.1	-0.2	-1.4	4.7	-4.3	21.4	0.9	0.0	0.1	1.2	1.3	96.0	1.4	0.0	0.0	0.5	0.4
PAR	0.050	3.7	16.4	18.9	-19.4	-10.9	0.4	-12.6	1.4	5.3	13.2	5.1	0.0	12.0	17.9	26.4	27.8	8.3	0.0	11.7
PRO	0.054	4.0	25.4	17.4	13.6	-6.7	2.4	-7.5	3.7	4.9	6.9	2.1	0.4	4.6	47.6	22.5	13.6	3.3	6.4	4.2
REC	0.045	8.3	-33.3	-10.3	4.1	0.0	-0.1	-7.6	11.5	1.3	0.3	0.0	0.0	3.9	84.0	3.3	0.5	0.0	0.0	1.7
REG	0.045	6.4	30.3	-29.1	7.3	-25.8	-0.2	6.0	4.4	10.7	1.7	25.3	0.0	2.5	35.1	30.6	2.1	25.4	0.0	1.4
REM	0.048	4.9	26.3	-27.3	3.7	13.0	1.5	4.4	3.3	10.6	0.5	6.9	0.1	1.4	37.3	40.1	0.7	9.2	0.1	1.1
SEN	0.054	3.8	16.0	22.5	21.2	0.5	4.0	-2.6	1.5	0.1	16.9	0.0	1.0	0.5	19.8	39.3	34.7	0.0	1.2	0.5
SEX	0.047	6.4	31.1	-29.2	11.2	-17.8	0.7	5.0	5.5	12.0	4.1	12.7	0.0	1.8	44.1	34.4	5.0	12.7	0.0	1.0
SOU	0.051	3.7	22.5	-18.7	-5.8	13.5	-6.0	-8.7	2.8	3.3	1.2	7.9	2.2	5.8	38.4	26.5	2.5	13.9	2.7	5.7
SUP	0.046	4.5	14.2	14.8	-31.4	-14.6	-6.1	-4.0	1.0	3.0	31.7	8.3	2.1	1.1	11.3	12.3	55.2	12.0	2.1	0.9

Tableau 3 : Tableau des contributions

- <u>Interprétation des résultats à partir du graphique</u> (graphe 4)

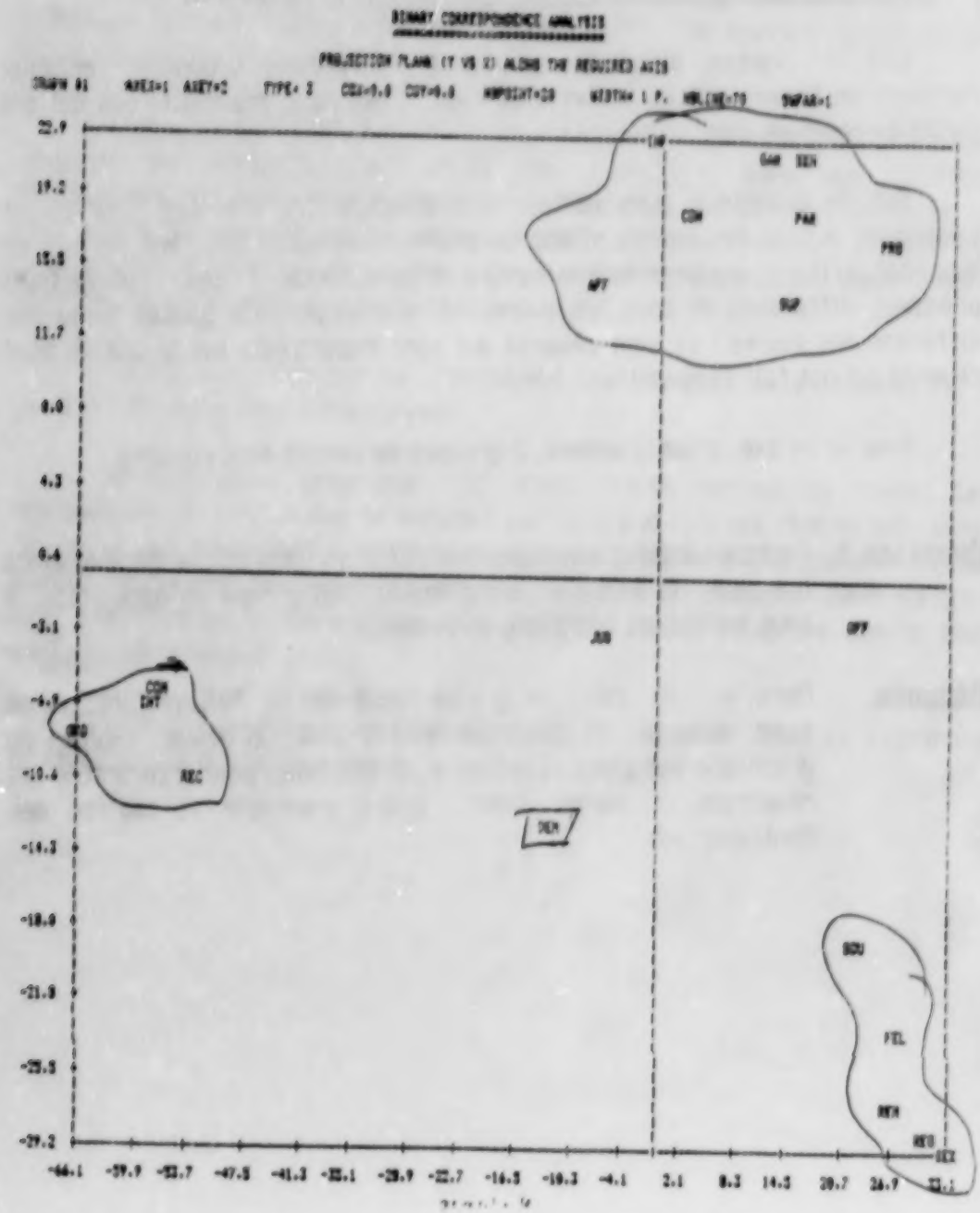
Les 4 verbes directifs commander, interdire, ordonner, réclamer forment, on l'a vu, 70% de l'inertie de l'axe, c'est donc vraiment eux qui ont formé ce premier axe.

Sur le graphique, ces verbes sont proches les uns des autres et s'opposent à tous les autres. Dans un premier temps il est donc permis de supposer qu'il y a une première structure très importante : ces 4 verbes **très proches**, différents de tous les autres. Il y a un premier paquet homogène différent des autres : ce sont ceux-là qui sont importants, parce que ce sont ceux-là qui ont fait l'opposition.

Pour le 2e axe, graphiquement, 2 groupes de verbes sont opposés.

<u>Question 3</u>: Pourquoi mettre en opposition ici 2 groupes sur le 2e axe, alors que pour le 1er axe, simplement un groupe en opposition à tous les autres était mis en évidence?

<u>Réponse</u>: Dans le 1er cas, ce groupe représentait 70% de l'inertie de l'axe, dans le 2e cas, seulement 30%, un nouvel examen du graphique est alors nécessaire. Il y a donc, pour interpréter les résultats, un va-et- vient entre graphique et tableau des contributions.


BINARY CORRESPONDENCE ANALYSIS
PROJECTION PLANE (Y VS X) ALONG THE REQUIRED AXIS
22.9
19.2
15.5
11.7
8.0
4.3
0.6
-3.1
-6.8
-10.6
-14.3
-18.0
-21.8
-25.5
-29.2
-44.1
-39.9
-33.7
-27.5
-21.3
-15.1
-8.9
-2.7
3.5
9.7
15.9
22.1
28.3
34.5
40.7
46.9
53.1
59.3
65.5
71.7
77.9
84.1
90.3
96.5
102.7
108.9
115.1
121.3
127.5
133.7
139.9
146.1
152.3
158.5
164.7
170.9
177.1
183.3
189.5
195.7
201.9
208.1
214.3
220.5
226.7
232.9
239.1
245.3

L'étude du graphique, permet de distinguer nettement 3 paquets. Le graphique fait aussi apparaître des verbes isolés, par exemple demander qui est isolé des 4 verbes directifs.

Une des chercheures, Sylvie Cantin, responsable de cette recherche, fait observer que demander peut aussi bien s'employer dans une situation hiérarchique que dans une situation non-hiérarchique, ce qui peut expliquer sa position isolée par rapport aux verbes directifs parce que son côté directif est plus atténué par rapport aux autres. Demander est utilisé autant dans le sens d' exiger que de demander une permission.

A droite du graphique se retrouvent les verbes plus coopératifs, à gauche les verbes plus autoritaires.

L'interprétation graphique ici s'est faite surtout au niveau des regroupements. Plutôt que de nommer par un grand concept chacun des axes, la recherche consistait à voir si les verbes d'une même catégorie théorique étaient regroupés. La structure est une structure en paquets. Dans d'autres cas, l'interprétation sera différente, lorsqu'on a des axes de points plus continus par exemple.

D'autres exemples seront examinés dans l'atelier. Nous ne reprenons brièvement ici que le 2e exemple.

II / EXEMPLE 2

Dans ce cas, il s'agit d'une analyse de correspondances multiples. Le tableau des données est un tableau disjonctif complet.

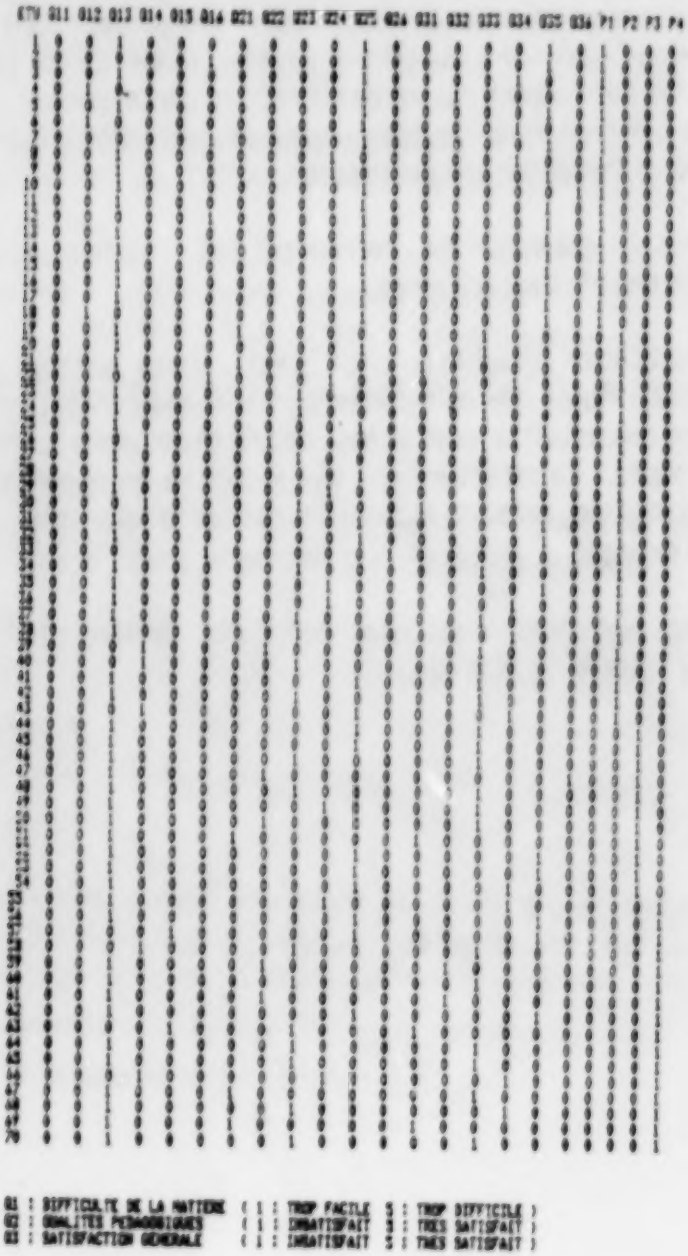


Tableau 1

Ces résultats sont tirés d'un questionnaire d'évaluation des enseignements passé à la fin de cours universitaires (il ne s'agit nullement, dans ce cas, d'une recherche). 70 étudiants ont répondu à ce questionnaire, pour 4 professeurs différents : P_1, P_2, P_3, P_4. 3 types de questions sont ici proposées: difficultés de la matière (Q_1), qualités pédagogiques (Q_2), satisfaction générale (Q_3).

Pour chacune des questions, une échelle de 1 à 5 permet de mesurer le degré de satisfaction de l'étudiant face au cours (6 correspondant aux non-réponses).

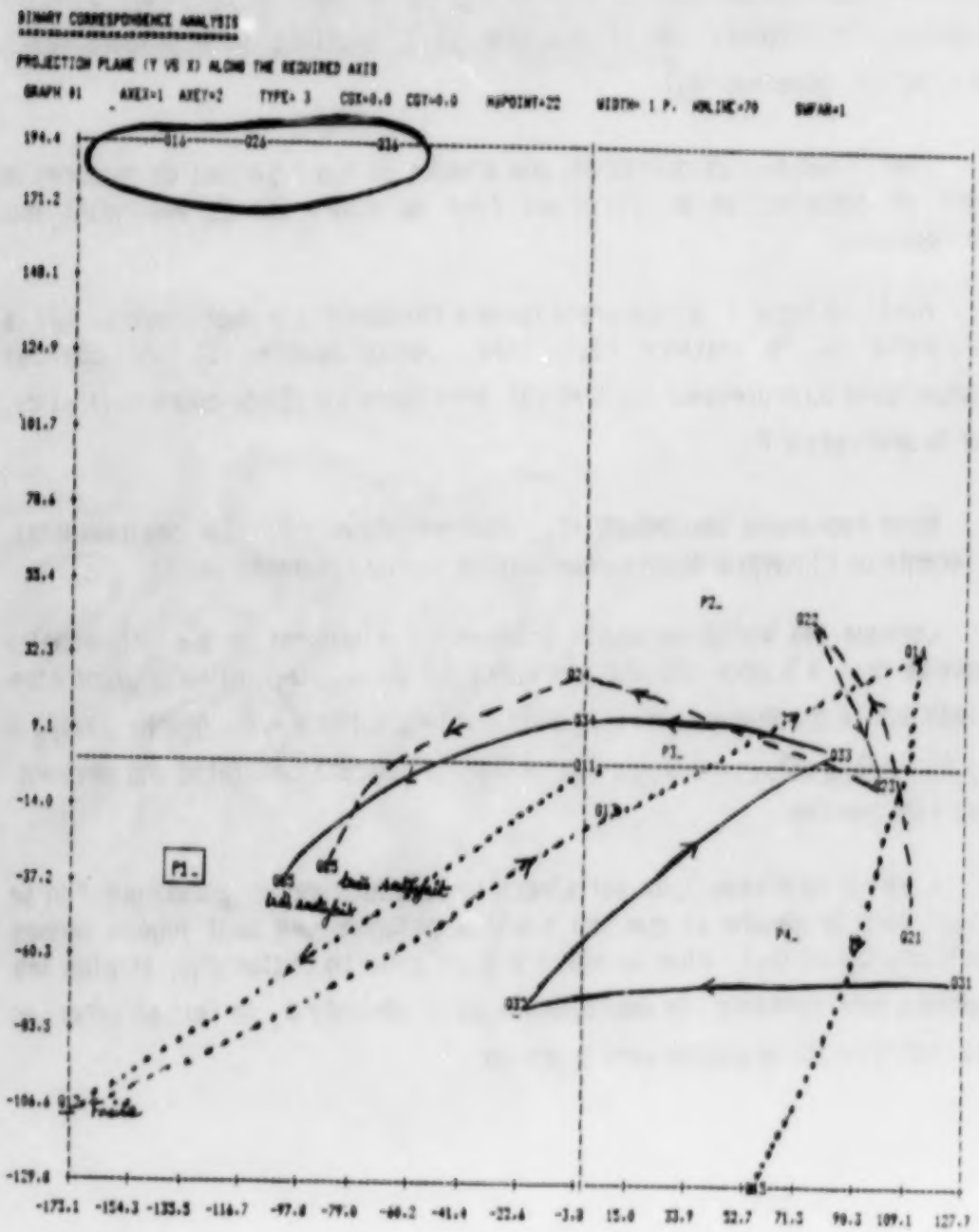
Ainsi, la ligne 1 s'interprète comme l'étudiant 1 a jugé moyenne (3) la complexité de la matière (Q_1), très satisfaisante (5) les qualités pédagogiques du professeur (Q_2), et est très satisfait (5) du cours (Q_3), et ce, pour le professeur P_1.

Nous reprenons seulement ici l'interprétation graphique des résultats, différente de l'interprétation menée dans le cas du 1er exemple.

Lorsque les variables sont ordonnées (ce qui est le cas ici, échelle ordonnée de 1 à 5 pour chacune des variables Q_1, Q_2, Q_3), elles peuvent être reliées sur le graphique ; on relie donc sur le graphique (voir graphe 2) Q_{11} à Q_{12} à ... à Q_{15} ; Q_{21} ... à Q_{25} ; Q_{31}...à Q_{35}. Plusieurs constatations peuvent alors être émises:

A partir du niveau 3, la satisfaction est d'autant plus grande que l'on se dirige vers la gauche et que les qualités pédagogiques sont jugées bonnes (questions Q_2 et Q_3). Plus la matière est facile (question Q_1), et plus les étudiants sont contents : le déplacement de la variable Q_1 se fait en effet en sens contraire, de la gauche vers la droite.

Seul le professeur P_1 se retrouve du côté gauche, côté correspondant à un taux de satisfaction plus élevé.



CONCLUSION

Les exemples très simples repris ici en atelier ont permis de vivre une démarche d'analyse de données, en en parcourant toutes les étapes. Ils nous ont ainsi permis de mieux comprende la méthode d'analyse des correspondances, en en cernant davantage la portée.

ATELIER 2 : ANALYSE DE DONNEES DIDACTIQUES
(méthode classificatoire)

Personne-ressource : Régis Gras

L'atelier est organisé autour de données brutes provenant de chercheures du CIRADE. Un programme créé par Régis Gras, adapté pour fonctionner sur le micro-ordinateur IBM-PC, nous permet de traiter les données sur place. Celles-ci seront ainsi examinées et analysées avec les participants de l'atelier.

Au préalable, un traitement mathématique local (calculs faits sur place) servira de retour à la méthode classificatoire et d'exercice d'acquisition de celle-ci.

I/ CALCULS DES INDICES DE SIMILARITE, OU COMMENT EST CONSTRUIT L'ARBRE DE CLASSIFICATION

Un cas précis est présenté (cas fictif, car il n'y a aucun intérêt à faire une analyse sur un aussi petit nombre de variables). L'exemple vise seulement à mieux comprendre la méthode.

	a	b	c	d
1	1	1	0	1
2	0	1	0	1
3	1	0	1	0

3 individus et 4 variables constituent l'exemple.

L'examen du tableau montre que les deux variables b et d sont celles qui se ressemblent le plus. En théorie, dans la méthode d'analyse classificatoire, ces deux variables devraient donc être réunies.

- Un court rappel sur le calcul de l'indice de similarité entre 2 variables est opportun, avant de passer à l'analyse proprement dite.

Soit E l'ensemble des élèves, et prenons deux questions a et b par exemple.

Note : si deux ensembles sont petits, il est exceptionnel qu'il y ait intersection. L'intersection ici possède un certain effectif, et donc une similarité existe. Cette similarité doit ensuite être mesurée à l'aide d'un indice.

$E_{(b)}$ représente les élèves qui ont répondu à la question b.
$E_{(a)}$ représente ceux qui ont répondu à la question a.

La similarité de a et b est liée au cardinal de l'intersection de $E_{(a)}$ et $E_{(b)}$: $|E(a \cap b)|$.

On peut calculer $P_r [| X \wedge Y | = s]$ par loi hypergéométrique, soit approximativement par loi de Poisson P (n p(a) p(b)).

L'indice de similarité entre a et b est alors donné par :

$$s(a, b) = \frac{n(a \wedge b) - \frac{n(a)\, n(b)}{n}}{\sqrt{\frac{n(a)\, n(b)}{n}}}$$

(indice de similarité calculé par le programme).

On calcule alors $s'(a,b)$, indice gaussien.

- Première étape : calcul des indices de similarité entre deux variables

Pour l'exemple présenté ci-dessus, le calcul de similarité entre les différentes variables deux à deux constitue la première étape, avant de tenter de prédire ce que devrait être l'arbre de classification suite à ces calculs.

	a	b	c	d
1	1	1	0	1
2	0	1	0	1
3	1	0	1	0

$$s(b, d) = \frac{n(b \wedge d) - \frac{n(b)n(d)}{n}}{\sqrt{\frac{n(b)n(d)}{n}}}$$

$n(d \wedge b)$ = nombre de réussites communes à d et b (2);
$n(b)$ = nombre de réussites en b (2);
$n(d)$ = nombre de réussites en d (2);
n = nombre total (3).

donc $$s(b,d) = \frac{2 - \frac{2 \times 2}{3}}{\sqrt{4/3}} = \frac{2/3}{2/\sqrt{3}} = \frac{\sqrt{3}}{3} \simeq 0{,}57$$

De la même façon, on peut calculer

$$s(a,c) = \frac{1 - \frac{2 \times 1}{3}}{\sqrt{2/3}} = \frac{1/2}{\frac{\sqrt{2}}{\sqrt{3}}} = \frac{\sqrt{3}}{3\sqrt{2}} = 0{,}4$$

Donc b ressemble plus à d, que a ne ressemble à c (ce qui dans ce cas fictif,

petit nombre d'individus, se voit effectivement sur le tableau)

L'indice gaussien (calculé par le programme) correspond à

$s'(b,d) = 0,725$
$s'(a,c) = 0,675$
$s'(b,c) = s'(d,c) = 0,25$
$s'(a,b) = s'(a,d) = 0,4$

- A partir de ces calculs, prédiction de comment la classification va se faire?

Quelles sont les variables qui se réunissent en premier?

L'indice de similarité le plus fort est s'(b, d), par conséquent au 1er niveau, les variables qui vont se réunir sont { b, d }.

Les variables sont donc maintenant réduites à : { a, { b,d, } c } et il est nécessaire de calculer l'indice de similarité entre cette fois-ci une variable et une classe. Comment le calculer ? L'indice de similarité entre deux classes se calculera de la même façon).

$s(a, \{ b,d \}) = \max (s(a,b), s(a,d))$, le maximum des indices de similarité entre les variables prises deux à deux est retenu. Mais le fait de prendre le maximum peut en fait fragmenter la ressemblance. Il y a un moyen alors d'atténuer cet effet de maximum, en élevant à la puissance deux (l'indice de similitude étant plus petit que 1 , le carré du nombre devient plus petit).

$$\text{Ainsi } s(a, \{ b, d \}) = [\max (s(a, b), s(a, d))]^2$$

$$= [\max (0,4) ; (0,4)]^2 = 0,16$$

Note: Il est nécessaire d'élever à la puissance 3 s'il y a 3 indices de similarité en jeu; le maximun est pris n fois s'il y a n indices de similarité en jeu.

Les autres indices de similarité sont

$s(a,c) = 0,675$

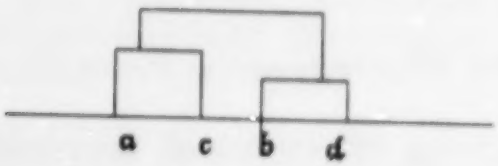
et $s(c,\{ b,d( ) = [\max (s(c,b), s(c,d))]^2 = (0,25)^2 = \frac{1}{16} = 0,06$

Par conséquent, au 2e niveau, les variables qui vont se réunir sont a et c.

L'arbre de classification dans ce cas sera donc :

1er niveau {b,d}
2e niveau {a,c}
3e niveau {(a,c) , (b,d)}

et le graphique correspondant devrait donc être (arbre de classification)



II/ EXEMPLE DE DONNEES BRUTES TRAITEES SUR PLACE

- L'exemple

Un exemple réel de recherche menée au CIRADE et de données brutes partielles tirées d'une expérimentation qui vient d'être effectuée fait ici l'objet d'une analyse.

L'expérience qui a été menée concerne deux populations d'enfants (un groupe expérimental et un groupe témoin). Elle fait suite à un travail de recherche mené pendant trois ans par Nadine Bednarz et Bernadette Janvier à partir d'une réflexion sur l'apprentissage de la numération et des opérations au primaire[1]. Lors de ce travail, ces deux chercheures ont pris en charge l'enseignement des mathématiques auprès d'un même groupe d'enfants (groupe expérimental) de la 1ère à la 3e année (6-7 à 8-9 ans), développant une stratégie didactique basée sur une approche constructiviste du concept de numération et de son apprentissage. Les données présentées ici correspondent à une évaluation <u>après 3 ans</u> de l'impact de cette intervention. Des 23 enfants de départ, il ne reste maintenant que 16 enfants (actuellement en 6e année) qui ont été confrontés pendant 3 ans, après cette intervention, à un enseignement régulier des Mathématiques.

1 - voir, entre autres, deux articles parus dans IN, novembre et décembre 84.

L'évaluation s'est faite en deux temps à l'aide d'instruments élaborés à partir d'un cadre de référence:

- un test écrit, dont les données (partielles) présentées dans cet atelier sont tirées.

- une entrevue individuelle avec chaque enfant.

Ces épreuves ont été passées auprés des 16 enfants ayant participé à l'étude de 1980 à1983 (groupe expérimental) et auprès de 19 enfants de 6e année de la même école ayant vécu un enseignement régulier de l'arithmétique (groupe témoin).

Le test écrit portait surtout sur les opérations (sens), beaucoup travaillés avec les enfants à l'époque, et sur les décimaux (auxquels ils ont été sensibilisés via l'approche constructiviste de la numération..).

Quelques items ont été choisis, pour l'atelier, parmi l'ensemble de ceux auxquels l'enfant a été soumis (le nombre étant trop grand pour les traiter tous). Rappelons que le but de l'atelier n'était pas de traiter véritablement de toutes ces données, mais de mieux comprendre, via cet exemple, la méthode de classification. Les items sélectionnés apparaissent en annexe.

Les données par groupe (groupe expérimental et groupe témoin) sont rentrées sur ordinateur à l'aide d'un tableau du type:

	Items du test	
Enfants (quelques uns)		données entrées codées réussite, échec : Notées 1 (échec), 0 (réussite), les ressemblances au niveau des échecs faisant ici l'objet de l'analyse (la même chose sera faite dans la reprise de l'atelier au niveau des réussites avec un autre groupe).

Remarque : Dans cette étude ce sont en fait les ressemblances et différences au niveau des items entre chacun des groupes qui sont étudiées. Le fait de rentrer les données par réussite ou échec ne correspond donc peut-être pas à nos attentes, puisqu'en fait ce sont avant tout les stratégies qu'il est intéressant d'étudier.

Au préalable des participants de l'atelier font des prédictions de classement sur les résultats des items qu'ils ont entre les mains.

Les prédictions suivantes sont avancées :

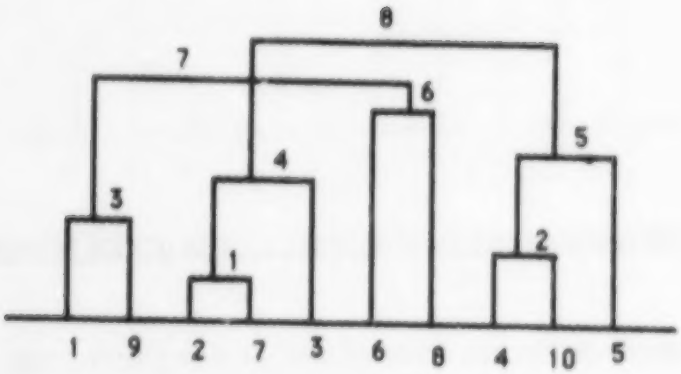
items 3, 4, 5	(puisque les trois touchent au contenu de la division);
items 1, 6, 7	(les trois faisant appel dans la tâche à une certaine initiative);
item 10 ,	l'item 10 est isolé des autres (contenu différent...);
items 2, 8, 9	(touchent tous trois à l'item calculatrice);
items 6, 7, 8, 9 10	

Les résultats (traitement par l'ordinateur), pour le groupe expérimental (quelques enfants), sont les suivants :

1er niveau (2,7)
2e niveau (4, 10)
3e niveau (1,9)
4e niveau ((2,7), 3)
5e niveau ((4,10),5)
6e niveau (6,8)

	3e niveau	4e niveau	6e niveau
7e niveau	((1,9) ,	((2,7),3) ,	(6,8))
8e niveau	[((1,9) ,	((2,7),3) ,	(6,8)), ((4,10), 5)]
	7e niveau		5e niveau

Le graphique de l'arbre est donc:



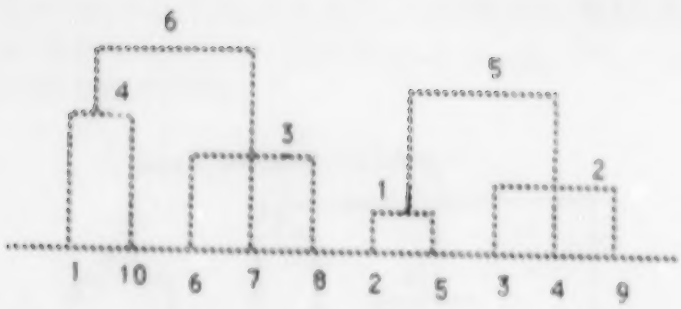
Arbre hiérarchique des 10 problèmes pour le groupe expérimental

Une première analyse de cet arbre fait apparaître 2 classes séparées 4,10,5 d'une part, et les autres (items qui se rejoignent seulement au 8e niveau).

Les regroupements qui apparaissent ici doivent être interprétés par le didacticien. Ainsi 2 et 7 apparaissent regroupés au 1er niveau. Dans ces deux items, on observe le recours à une conception inappropriée de la division développée chez les enfants "on ne peut diviser un petit nombre par un plus grand nombre".

- Le même travail est repris maintenant pour quelques enfants de l'autre groupe (groupe témoin), les résultats qui apparaissent alors sont les suivants : (classification différente de celle du premier groupe).

1er niveau	(2,5)
2e niveau	(3,4,9)
3e niveau	(6,7,8)
4e niveau	((2,5), (3,4,9))
5e niveau	(1,10)
6e niveau	((1,10), (6,7,8))



Arbre hiérarchique des 10 problèmes pour le groupe témoin

Deux classes distinctes apparaissent dans le graphe final:
(2, 5, 3, 4, 9) et (1, 10, 6, 7, 8).

Remarques: Des variantes seront essayées dans la reprise de l'atelier, montrant comment on pourrait changer la façon de rentrer les données et à quels types de résultats cela conduirait. Ainsi,

- les mêmes données seront rentrées en codant cette fois-ci 1 pour les réussites, 0 pour les échecs, faisant alors apparaître les ressemblances au niveau des réussites.

- on reprendra les mêmes problèmes en mettant les élèves comme variables.

	Enfants
Problèmes	

Les élèves du groupe témoin et du groupe expérimental seront alors classifiés selon leurs profils, afin de voir si les 2 classes obtenues lors de la classification seront d'une part les élèves du groupe témoin et d'autre part les élèves du groupe expérimental.

Annexe (Atelier 2)

-1-

Le professeur d'arts prépare des morceaux de feutrine pour un bricolage. Il a 47 morceaux de feutrine comme celle-ci qu'il veut partager également.

Pour savoir ce qu'il donnera à chacun des enfants de sa classe, il calcule sur sa calculatrice qui lui dit de donner à chacun :

Quelle réponse le professeur a eu sur sa calculatrice ?

-2-

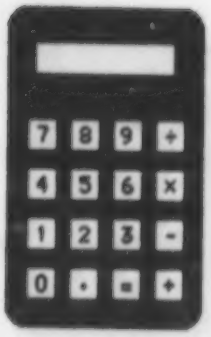
2- Regarde ce que j'ai fait sur la calculatrice

Qu'est-ce que la calculatrice va afficher?

-3-

3- Pour mon voilier, j'ai besoin de plusieurs bouts de corde de 0.26 mètre de long. Ecris ce que tu ferais sur la calculatrice pour trouver combien de bouts de corde je peux préparer avec 1.96 mètre de corde?

Je dis ce que je fais sur la calculatrice:



-4-

4- Jean remplit un contenant d'eau de 17 litres avec un petit pot contenant 0.2 litre. **Montre-moi comment tu ferais** pour trouver combien de fois il doit utiliser le petit pot pour remplir son contenant de 17 litres?

Débrouille-toi, je veux voir ce que tu fais:

-5-

5- Alain parcourt la longueur du corridor de 36 mètres en faisant des pas égaux. Il fait 32 pas et arrive au bout du corridor. Montre-moi comment tu ferais pour trouver la longueur de son pas.

Débrouille-toi, je veux voir ce que tu fais:

-6-

6- Le nombre 253.351 est multiplié par 100.

Le 1 qu'on obtient dans le résultat a la même valeur que le 1 dans lequel des nombres suivants? Entoure-le.

426.72100 2432.125 4267.510 3571.52

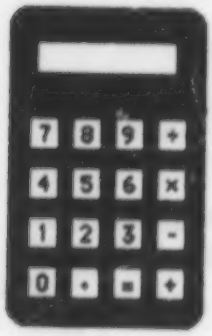
-7-

7- 25 enfants se sont entendus pour nettoyer le terrain de football ensemble, ça leur a rapporté 22.00$. Combien chacun reçoit-il?

Jean dit qu'il faut faire 25 - 22 sur la calculatrice pour trouver la réponse.

Jean a-t-il raison ou tort?_____

S'il a tort, dis ce qu'il devrait faire:



-8-

8- Sur quelles touches de la calculatrice appuierais-tu pour calculer trois plus onze dixièmes?

-9-

9- Sur quelles touches de la calculatrice appuierais-tu pour calculer deux fois onze millièmes?

-10-

10- J'ai deux nombres 0.86 et 0.752. Jean dit qu'il peut trouver un nombre entre les deux. Paul dit que non. Qui a raison? Explique.

COMPTE RENDU DE LA TABLE-RONDE
autour de questions soulevées par les recherches actuelles sur l'apprentissage en regard des méthodes d'analyse de données présentées lors de ce colloque

Intervenants :

Régis Gras	(Institut de recherche mathématique de Rennes)
Brigitte Escofier	(Institut de recherche en informatique et Systèmes aléatoires de Rennes)
Jean-Marie Van der Maren	(Faculté des Sciences de l'éducation, Université de Montréal)

Nous reprenons dans ce compte rendu les principales questions formulées par différents participants, et les réponses et discussions qu'elles ont suscitées.

•Question I relative aux études de type longitudinal: comment, par les méthodes proposées, les populations enregistrées à des moments différents peuvent-elles être comparées?

Dans une recherche classique de type pré-test, post-test, la performance des étudiants est mesurée. Mais jusqu'à quel point la différence obtenue dans les résultats nous renseigne-t-elle sur leur évolution.

Dans une analyse de correspondances, un moment donné de l'apprentissage correspond à un certain nuage de points, un certain état de performance dans lequel les étudiants sont répartis (n stratégies différentes). A un moment ultérieur, l'analyse donne un autre nuage de points.

La transformation des nuages est donc l'objet d'intérêt: est-ce que cette transformation est telle que l'apprentissage peut être considéré comme significatif, en quoi l'est-il, selon quelles caractéristiques?

Comment les méthodes proposées dans ce colloque peuvent-elles permettre de clarifier cette question, de quantifier le passage d'une stratégie à l'autre...

- **Réponse de Mme Escofier (pour l'analyse des correspondances):**

Au départ,

L'étude porte sur un certain nombre d'étudiants et un certain nombre d'épreuves qui leur ont été proposées. Les résultats de ces épreuves donnent des résultats identifiés en termes de stratégies (stratégies conduisant à des échecs, stratégies vers le haut conduisant à des réussites...).

Un tableau de données du type suivant peut être construit:

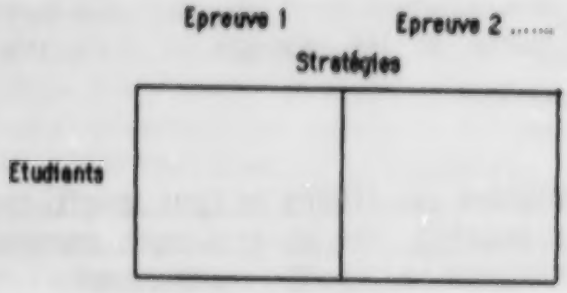


Tableau 1

Une analyse de correspondances réalisée à partir de ce tableau permettra probablement de dégager un axe caractérisé par le déplacement des étudiants d'une stratégie à un autre stratégie, d'une stratégie plus naïve à quelque chose de plus complexe.

Ceci ne règle toutefois pas la question de comparer ce qui se passe avant et après l'apprentissage.

• Dans le cas d'une même série d'épreuves passée après l'apprentissage, un tableau de données, retraçant les étudiants avant l'apprentissage et les mêmes étudiants après l'apprentissage, peut être défini.

	Epreuves Stratégies
Etudiants avant l'apprentissage	
Etudiants après l'apprentissage	

Tableau 2

- Une analyse des correspondance est réalisée à partir de ce tableau et permet l'examen des nuages de points : chaque individu y est représenté avant et après l'apprentissage.

- Cette analyse permet probablement de retrouver un axe caractérisé par un déplacement d'une "mauvaise" stratégie à une stratégie plus évoluée. Elle permet également de voir comment les individus se situent par rapport à cet axe, en les regardant un à un (s'il n'y en a pas trop), ou tous ensemble.

- La moyenne des individus pourra également être calculée avant et après l'apprentissage.

En fait, si on veut comparer la moyenne avant et après, il suffit d'ajouter au tableau initial de données (tableau 2) deux lignes : la somme des lignes des étudiants avant l'apprentissage, la somme des lignes des étudiants après.

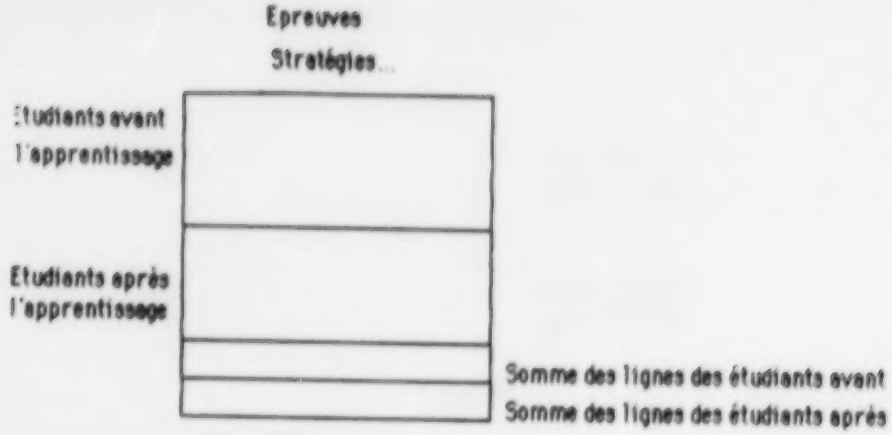


Tableau 3

La moyenne des étudiants avant l'apprentissage et la moyenne des étudiants après l'apprentissage apparaîtra automatiquement sur le graphique.

Donc l'analyse des résultats (référence est faite ici au graphique) qui sortira de cette analyse des correspondances nous permettra de visualiser la moyenne des étudiants avant (représentée par un point) et la moyenne des étudiants après l'apprentissage; de voir s'il y a eu évolution et dans quel sens (par exemple évolution des étudiants qui se trouvent en haut et en bas de l'axe) et, de plus, de voir cette évolution étudiant par étudiant.

Eventuellement, s'il y a plusieurs stratégies d'évolution, elles pourront être identifiées, et pour les individus apparaissant à 2 endroits, ce qui est figuré par les points, il sera possible de regarder comment ils se sont déplacés.

C'est un peu comme si deux nuages étaient comparés par un nouveau nuage.

Il suffit de regarder le nuage d'individus avant et le nuage d'individus après, ainsi que le sens (l'allure) de l'évolution.

• Ou bien il est facile de voir qu'il y a eu évolution :

Par exemple

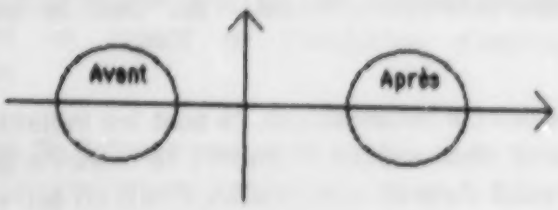


Figure 1

• Ou bien les nuages sont entremêlés :

Par exemple

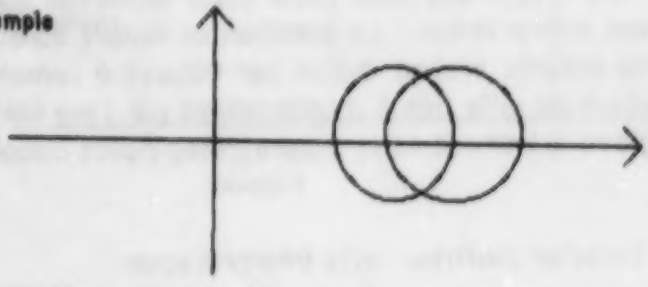


Figure 2

si les nuages se présentent de cette façon (fig. 2) il est intéressant de regarder les individus;

si un individu est permuté, on peut essayer d'expliquer pourquoi cet individu n'a pas agi comme les autres;

si les nuages sont très entremêlés, il est nécessaire d'expliquer pourquoi ils le sont : est-ce parce que les individus vont tous dans le même sens alors que l'évolution des individus est trop petite par rapport à l'ensemble; ou bien parce qu'il y a eu à la fois évolution d'individus dans un sens et dans l'autre.

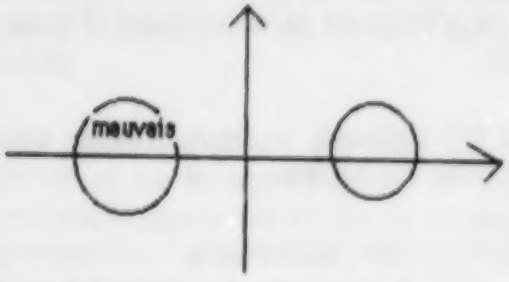
Intérêt de cette technique par rapport à l'analyse multivariée.

Dans le cas de l'analyse multivariée, l'analyse est multidimensionnelle; dans l'analyse des correspondances, il y a une seule variable. L'analyse et l'interprétation peuvent donc porter effectivement sur le contenu dans l'analyse des correspondances, ce qui n'est pas le cas pour l'analyse multivariée.

Dans l'analyse des correspondances, ce sont les individus qui sont d'abord observés afin de voir chaque petit élément : la moyenne peut être la même, mais certains ont évolué dans un sens, d'autres dans un autre sens; ce qui n'est pas le cas dans l'analyse multivariée, parce que tous les individus sont présents. L'analyse de correspondances est donc presque une analyse de contenu.

Régis Gras revient sur la question de départ et explique que Brigitte Escofier a fait l'hypothèse dans toute cette démarche que le premier facteur avait un sens interprétable. La question de départ consistait surtout à voir comment les enfants avaient évolué par rapport à l'ensemble des stratégies. Cette évolution est-elle liée à un glissement sur l'axe (défini comme opposant une stratégie relativement naïve à une stratégie plus élaborée)?

Mme Escofier confirme cette interprétation.



Si des individus sont très mauvais (stratégies relativement naïves) et que ceux-ci ont beaucoup bougé après l'apprentissage et les autres peu, le graphique le montrera.

Régis Gras fait observer que, par contre, le nuage de stratégies ne va pas changer, il reste projeté. Le glissement va se faire par rapport aux stratégies identifiées sur le dessin. Donc le mouvement doit se faire dans le sens de certaines stratégies ou de reculs, toujours par rapport aux deux ensembles.

Brigitte Escofier ajoute ici un complément de réponse sur l'analyse des correspondances en regard de l'évolution d'individus avant et après expérimentation.

Le tableau de données peut être construit en juxtaposant les étudiants avant et après l'apprentissage (tableau 3). Mais le tableau de données pourrait aussi être construit en juxtaposant les épreuves du premier test et celles du second test pour les mêmes étudiants.

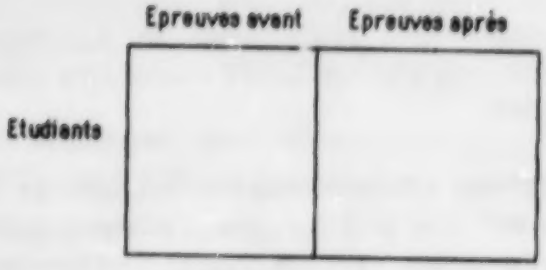


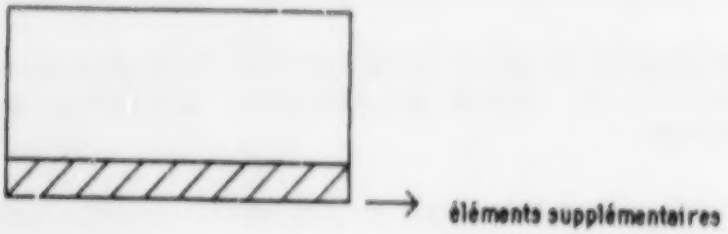
Tableau 4

L'analyse conduit alors, non plus à voir la trajectoire de chacun des étudiants, mais à voir mieux les relations entre les résultats du pré-test et du post test. Ceci permet par ailleurs d'avoir des épreuves différentes avant et après expérimentation. Une analyse n'exclut pas l'autre, les deux peuvent être faites, suivant l'intérêt.

En faisant des classifications, une matrice est construite où les graphes obtenus avant et après apparaissent, ce qui permet d'étudier la corrélation entre les graphiques avant et après, et nous permet de voir qui a changé. Est-ce les sujets qui apparaissent faibles dans le graphique avant l'apprentissage, qui se sont déplacés vers la partie graphique "forte" ou pas? Différentes possibilités de ce genre peuvent être exploitées.

Il est possible de mettre en supplémentaire également certaines variables. Mme Escofier précise ce que cela signifie.

Dans le tableau de données initial, il est possible de mettre une variable en supplémentaire.



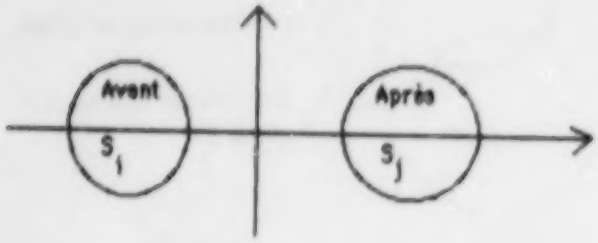
Par exemple, avec deux populations, il peut être intéressant de connaître immédiatement les moyennes de chacune de ces populations sans que celles-ci influent sur la formation des axes. Des lignes qui correspondent à la moyenne de chaque population sont alors mises en supplémentaire. Ces variables n'influenceront pas l'analyse, par contre il peut être intéressant de regarder où elles sont projetées.

Cette technique, utilisée fréquemment dans la méthode d'analyse des correspondances est très pratique pour l'interprétation, pour savoir où se retrouvent certains points, sans que ceux-ci n'influencent l'analyse, la mise en évidence de facteurs principaux.

- **Réponse de Régis Gras, (pour la méthode classificatoire)** :

Le problème est de restituer la dynamique de changement d'une population avant et d'une population après. Nous sommes intéressés à voir avant tout si des étudiants qui utilisaient une stratégie S_i, utilisent maintenant prioritairement une stratégie S_j. Les méthodes d'analyse de données didactiques exposées dans ce colloque (méthodes classificatoire et implicative) ne sont pas multidimensionnelles, comme l'est l'analyse des correspondances. Ces différentes analyses ne sont toutefois pas exclusives, elles se situent en complémentarité. La méthode classificatoire pourrait par exemple s'appuyer sur une hypothèse qui se serait dégagée d'une analyse des correspondances; celle-ci ayant permis de faire un premier tri sur les stratégies qui sont en filiation (stratégie antécédante, stratégie postérieure).

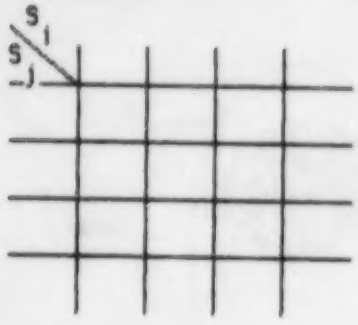
Le glissement de nuages d'individus dans l'analyse des correspondances permet de mettre en évidence cette filiation.



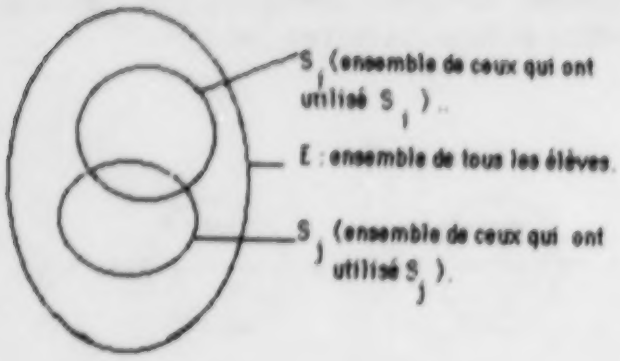
Les étudiants étaient au début centrés autour d'une stratégie S_i.

Maintenant, ils sont plus centrés autour d'une stratégie S_j.

Un tableau, qui étudie plus spécifiquement l'implication éventuelle entre des stratégies S_i et S_j, peut être réalisé en côtant par 1 et 0 la présence ou l'absence de la stratégie à un moment donné.



Le tableau peut alors être rempli comme le font les psychologues, ils calculent les marges...calculent le χ^2 pour voir s'il y a dépendance ou non... et l'étude de la mutation éventuelle de S_i à S_j peut suivre.



Quels sont les glissements de S_i à S_j? (S_j pouvant être de S'_i, i.e. les mêmes stratégies mais après expérimentation).

Si S_i et $S'i$, sont différenciés, c'est pour mieux voir le glissement, autrement dit derrière S'_i apparaît la première stratégie S_i mais cette stratégie est identifiée après un certain déroulement dans le temps (correspondant à un apprentissage).

Que signifie voir s'il y a un mouvement, une évolution de S_i vers S_j et mesurer cet écart par rapport à la méthode classificatoire ?

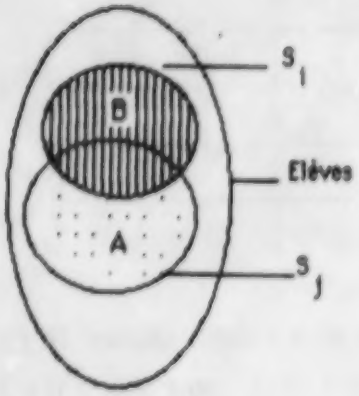


Figure 3.

Il est possible d'obtenir ceci (fig. 3) et de s'interroger alors sur ces résultats : est-il étonnant d'obtenir l'indépendance entre les deux types de graphiques A et B? ou bien il est possible d'étudier $S_i \rightarrow S_j$ ou $S_j \rightarrow S_i$.

Dans le cas où $S_i \Rightarrow S_j$, tous ceux qui ont utilisé la stratégie "i" utilisent également la stratégie "j".

Dans le cas où $S_j = S'_i$, l'assurance est donnée que chaque enfant qui utilisait une certaine stratégie S_i est resté, après l'expérimentation, utilisateur de la même stratégie S_i, ou a plutôt utilisé de préférence celle-ci.

Par des croisements deux à deux, il est ainsi possible par l'analyse classificatoire de revenir sur certains facteurs, après une étude relativement globale (l'analyse des correspondances est une analyse multidimensionnelle qui prend toute l'information et rentre toutes les stratégies).

Une typologie des stratégies (arbre de classification) pourrait être également construite, permettant de mettre en évidence celles qui ont des affinités.

- **Complément de réponse de Jean-Marie Van der Maren**

La question qui est posée est en fait celle du pré-test, post-test. Un des points importants, avec l'une ou l'autre de ces techniques, est de conserver les profils complets des sujets.

Avec une technique statistique habituelle, il est nécessaire de prévoir à l'avance tous les effets qui vont être examinés.

Mais, si dans le post-test de nouvelles stratégies apparaissent, cela peut être gênant. Avec une de ces techniques, il est possible de les incorporer dans le profil. Elles apparaîtront au post-test chez certains sujets, chez d'autres pas.

Ce problème est fréquent dans nos recherches centrées sur l'apprentissage, dans la mesure où lorsque l'enfant apprend, un changement de stratégies, un changement de représentations se crée.

Par conséquent, suivant le test qui aura été prévu, (pré-test), celui-ci n'est pas nécessairement adéquat après l'apprentissage puisqu'il y aura un changement de représentation. L'épreuve que nous avons construite est donc adéquate au pré-test ou au post-test, mais elle l'est rarement aux deux en

même temps. Alors en utilisant des stratégies d'analyse comme celles qui ont été présentées ici, le bassin des réponses possibles peut être augmenté ainsi que les modalités des réponses des sujets afin d'en tenir compte dans l'analyse, ce qui me semble beaucoup plus riche qu'une analyse statistique classique. Des analyses à postériori sont donc ici possibles sans devoir faire des prédictions au préalable sur ce qui va se passer.

•Question 2: Quelles différences et ressemblances existe-t-il entre ces techniques et le "cluster analysis" (analyse en grappes)?

- Réponse de Jean-Marie Van der Maren

Du point de vue de l'utilisateur, les analyses en grappes reviennent à faire de la classification. Il y a des classifications ascendantes, des classifications descendantes, il y a des classifications plus ou moins taxonomiques avec des possibilités d'inclusion de certains individus. Il y a toute une série de techniques qui existent, plusieurs personnes se sont lancées dans la construction d'indices qui permettent de regrouper les sujets. C'est un peu comme la technique de Lerman que Régis Gras a exposée dans ce colloque qui aboutit à la construction d'un graphe où peuvent être classés les individus.

Le "cluster analysis" aboutit à ce type de conclusions avec des techniques qui étaient peu sophistiquées dans les années 40, avec des techniques qui le sont maintenant plus et qui reprennent des indices de similarité comme ceux que Régis Gras a exposés. Il est aussi possible, après avoir fait une analyse en grappes, de faire une analyse en composantes principales et de voir comment les individus vont se placer sur les axes, après avoir vu comment ils se répartissent en grappes.

Brigitte Escofier souligne que ceci est fait systématiquement dans les bibliothèques d'analyse de données. Ainsi Addad* comporte des programmes de classification avec de très grosses données, les deux types d'analyse sont faites : conjointement avec l'analyse factorielle, des programmes de classification avec distances analogues sont utilisés afin de pouvoir comparer les résultats.

* Association française pour le développement et la diffusion de l'analyse des données.

•Question 3: Rapport entre le graphique obtenu (méthodes classificatoire et implicative) et l'arbre de la population qui était visé au départ?

Le problème est situé de la façon suivante: lorsque la construction d'un arbre est recherchée, derrière toute l'analyse il y a une population cible qui doit être déterminée et les observations obtenues proviennent d'un échantillon. Quand, dans la construction de l'arbre, il est décidé d'adjoindre une arête ou un sommet, un certain critère est utilisé ; dans la population réelle, l'arête n'existe pas; ceci est fait plusieurs fois, et l'arbre est finalement obtenu. Quel rapport existe-il entre cet arbre et l'arbre de la population visée au départ ?

- **Réponse de Jean-Marie Van der Maren**

Vous avez présupposé qu'il était possible de faire une inférence entre une population et un échantillon. Or rien ne nous dit actuellement que nous pouvons viser des lois générales et vouloir faire ces inférences. Peut-être que statistiquement et mathématiquement, ceci est un objet d'étude, mais comme psychologue, c'est un problème dans mon milieu qui commence à être de moins en moins important. Nous travaillons sur les problèmes de causalité, causalités qui sont locales et contextuelles, qui dépendent d'une espèce de pragmatisme de la situation et, le problème soulevé par la question nous apparaît de moins en moins important.

Le participant qui a posé la question revient sur celle-ci: le problème est tout de même important, si par exemple, je constate que dans l'échantillon de 15 étudiants il y a une amélioration, je suis intéressé à travailler sur une classe plus large que 15 individus.

Brigitte Escoffier souligne que cela dépend des domaines. Par exemple dans une enquête , ce ne sont pas les 1 500 individus interviewés qui sont intéressants, mais ces individus en tant qu'image de la population.

La démarche est la suivante, lorsque le nombre d'individus est suffisant. L'étude commence sur une sous-population d'individus. Par exemple, si l'échantillon comporte 1 000 individus, des analyses sur 800, 500 et 1 000 individus peuvent être faites successivement. Si les résultats sont les mêmes dans les trois cas, alors il est probable que le résultat sera cohérent. L'ajout de sujets ne fait pas beaucoup changer les nuages. Par la construction même,

les méthodes d'analyse factorielle sont assez stables.

Jean-Marie Van der Maren met en évidence une autre mentalité à mettre en action dans le cas d'études en psychologie, en éducation.

Pour une recherche en psychologie faite sur un nombre n d'individus (n petit), vouloir faire une généralisation à partir des résultats n'a pas de sens. Ce qui aurait du sens, c'est que le chercheur reprenne son expérience sur 12 individus, 12 autres... plutôt que de faire une généralisation à l'aide des statistiques. Il fait alors une abstraction, c'est-à-dire il essaie de trouver dans les principes qui règlent chacun de ces sous-groupes ce qui peut être commun. C'est une autre logique : la logique de l'inférence statistique est différente de la logique de l'abstration. Le psychologue, éducateur, étant donné le contexte dans lequel il travaille (pauvreté des populations, des sujets...), aurait intérêt à travailler par réplique, en tentant d'extraire ce qui est commun.

- **Régis Gras poursuit dans le même sens la réflexion.**

Ceux qui font des recherches en éducation n'écrivent jamais de théorèmes. Des hypothèses sont faites, les analyses proposées dans ce colloque livrent des hypothèses, la répétition, elle, permet de confronter ces hypothèses. Rien n'empêche de faire une analyse à l'aide des théorèmes relatifs au sondage qui permettrait de dire "J'ai trouvé par exemple 56% des étudiants sur les 400 qui ont eu tel comportement. Comme il y a 400 élèves la valeur à 95% est du type-----. Je peux identifier autour de 55% une certaine fourchette dans laquelle doivent se trouver à priori les vraies valeurs au seuil 5%. Il est possible de fonctionner ainsi, de sortir des arborescences (floues), une texture plus ou moins serrée, avec bien sûr des points d'interrogation qui demeurent.

La question de départ (Question 3) est de nouveau reprise : Idéalement, si une métrique entre les arborescences obtenues existait, cela permettrait de dire le type de structure qui est présenté, c'est à peu près telle chose, il serait possible d'englober d'autres arbres assez voisins de celui déjà donné et une meilleure vision découlerait de cette opération. Mais en fait ici dans l'analyse qui est faite c'est comme dire "le salaire moyen des gens c'est 25 000$, et ceux qui gagnent 3 000$ et ceux qui gagnent 100 000$ sont oubliés". Le problème demeure.

Jean-Marie Van der Maren soulève le fait qu'en recherche qualitative, il est possible de faire appel au groupe contrasté pour essayer de voir des choix rationnels (non libres), en fonction de certaines catégories de population. Le choix d'une population, et d'un échantillon ayant des caractéristiques inverses de ce qui nous paraît être des variables intéressantes nous permet de voir, à partir de ce contraste, comment se fait l'évolution. Ceci correspond, dans les méthodes anciennes, aux extrêmes dans un échantillon: les 25 moins forts, les 25 plus forts. Les contrastes peuvent également être multipliés, par exemple en faisant appel à un échantillon d'informateurs à qui on demande d'indiquer quels sont les gens qui ont une opinion inverse. A ces personnes, on demande également s'il y a une opinion contraire. De sorte qu'une représentation finit par se dessiner, qui n'est pas proportionnelle aux opinions qui sont émises, mais à la carte des opinions sans avoir leur pondération.

Cette question touche tout le problème, non pas de la taille de l'échantillon, mais de choses fondamentales qu'on oublie : la définition des caractéristiques spécifiques qui contrôlent un ensemble de variables à l'intérieur d'un échantillon, façon dont on peut garantir un maximum de fidélité dans le portrait obtenu. C'est la définition spécifique de "c'est quoi cet échantillon" qui s'applique surtout aux études sociologiques, pas tellement à la psychologie et à l'éducation.

•Question 4: Efficacité d'un modèle dans l'apprentissage; transfert des apprentissages?

Pour introduire cette question, l'exemple d'une recherche menée au CIRADE (mémoire de maîtrise sur l'enseignement des entiers relatifs) [1] est présentée.

---

1 Daïf Ahmed "Entiers relatifs: modèles et erreurs" mémoire de maîtrise dirigée par C. Janvier.

Pour enseigner un objet aussi abstrait que celui "d'entier relatif", problème qui se retrouve pour d'autres notions en mathématiques et en sciences, un modèle est souvent utilisé. La question de l'efficacité du modèle se pose alors.

Dans cette étude, deux groupes ont été comparés. Un groupe A (témoin) ayant vécu un apprentissage des "entiers relatifs" selon un modèle dit symbolique (un des modèles utilisé dans l'enseignement actuel), et un groupe B (expérimental) ayant vécu un apprentissage des "entiers relatifs", selon un modèle schématique. Après l'expérimentation, les étudiants des deux groupes ont été soumis à une épreuve où se retrouvaient des questions portant sur les "entiers relatifs" posées via une écriture dans le modèle d'une part, et posées via l'écriture conventionnelle (ex. $^{+}2 + {}^{-}5$) d'autre part. La question soulevée ici est donc celle de l'efficacité du modèle élaboré et de l'écriture intermédiaire dans ce modèle (modèle schématique): quel transfert à l'écriture conventionnelle y-a-t-il ?

Pour analyser les résultats et répondre à cette question, une multitude de tableaux du type suivant ont été utilisés.

M / A	Réussite	Echec
Réussite		
Echec		

A : écriture arithmétique conventionnelle.

M : écriture dans le modèle

Les performances dans le modèle et dans l'écriture conventionnelle sont comparées pour chaque item, et ce, pour les groupes témoin et expérimental. Plusieurs items (calculs sur les relatifs, problèmes) ont ainsi été analysés.

Deux questions fondamentales étaient sous-jacentes à cette étude.

- Est-ce que les étudiants utilisent vraiment l'écriture intermédiaire, le modèle ?

- Est-ce qu'il existe une différence fondamentale entre les deux modèles (symbolique et schématique) ?

Comment les méthodes d'analyse proposées dans ce colloque pourraient aider à répondre à cette question ?

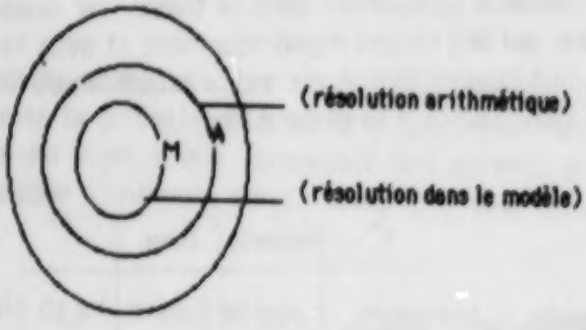
- **Réponse de Régis Gras**

Analyse de la situation (par la méthode implicative).
Il y a deux types de population :

- Groupe A (groupe témoin): groupe ayant suivi l'approche modèle symbolique ("modèle de la couleur")

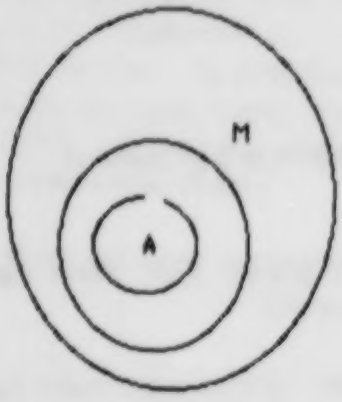
- Groupe B (groupe expérimental): groupe ayant suivi l'approche modèle schématique ("modèle des ballons")

On peut avoir cette situation, pour chaque item :



Tous ceux qui savent résoudre le problème posé dans les termes du modèle résolvent le problème posé en termes arithmétiques. Mais le passage par le modèle n'est pas nécessairement facilitateur pour résoudre le problème arithmétiquement, c'est peut-être les individus qui savent résoudre (avec leur compréhension spontanée des nombres relatifs), qui peuvent résoudre les problèmes lorsqu'ils sont posés dans les termes du modèle (ex. +3 -5 en termes de ballons qu'on rajoute, qu'on enlève---).

Ou alors l'investissement dans le modèle schématique (modèle du ballon) est un bon investissement



Ceux qui réussissent en termes algébriques sont parmi ceux qui réussissent les problèmes présentés en termes des ballons. Le modèle dans ce cas est facilitateur.

ou bien le modèle schématique est un obstacle pour se rendre directement au modèle formel.

L'étude des tableaux (présentés dans la thèse) par question, en regardant les effectifs de ceux qui ont réussi algébriquement et sous la forme du modèle, amène à constater la situation M → A, i.e. cette situation où ne peut être assuré que le modèle soit facilitateur à la résolution.

Ex. des tableaux

A \ M	Réussite	Echec	
Réussite	50	13	63
Echec	3	9	12
	53	22	75

<u>En calculant les coefficients d'implication</u> (méthode exposée au colloque) j'obtiens : A → M (0,90)
M → A (0,57).

L'examen des résultats (à partir des tableaux) nous amène à constater pour le groupe B, qu'il y a un temps où certaines représentations, certains modèles sont efficaces (c'est le cas du modèle des ballons pour des problèmes additifs), mais où ils peuvent par contre devenir des obstacles à la résolution d'autres problèmes (ex. : de problèmes mettant en jeu la soustraction ici). Ainsi, les étudiants, dans le cas d'un calcul à résoudre mettant en jeu trois moins successifs, réussissent mieux l'opération algébrique que l'opération posée dans les termes du modèle.

La situation est alors la suivante

Les résultats obtenus pour le groupe témoin (modèle symbolique) viennent confirmer la situation M ⇒ A; là le coefficient d'implication est très fort (0,999), ce qui montre que l'ensemble des enfants qui ont réussi dans le modèle est presque contenu dans l'ensemble des enfants qui ont réussi algébriquement.

Des calculs de χ^2 qui ont été faits montrent la dépendance. Ainsi, il y a des situations (items de calculs algébriques proposés aux individus ou problèmes) où c'est plutôt le modèle qui facilite et, il y a des situations où c'est plutôt le contraire.

Ceci est un moyen de répondre à la question qui était posée au départ : est-ce que le modèle est facilitateur ? On pourrait également construire un arbre en prenant toutes les variables et hiérarchiser alors différents items avec toute la combinatoire imaginable : addition de négatifs, soustraction de négatifs..... C'est possible de le faire, mais ce qui a été discuté ici plus spécifiquement c'est le côté transfert versus obstacle à la résolution algébrique.

- **Brigitte Escofier,** propose de regarder globalement, via l'analyse des correspondances, et non item par item, et d'observer alors les tendances qui se

dégagent d'un ensemble d'items.

Pour faire une telle analyse, un tableau de données est construit de la façon suivante: les individus sont placés sur les lignes et les items sur les colonnes, avec pour chaque item, 2 colonnes pour réussite - échec à l'item dans le modèle schématique; réussite - échec à l'item dans le modèle symbolique... et ceci pour toutes les questions.

L'analyse devrait montrer s'il y a continuité globale dans l'analyse entre échec dans un modèle et échec dans l'autre, suivant le modèle et les individus. Ce qui permettrait éventuellement de pouvoir avoir des effets cumulés, c'est-à-dire de voir ce qu'il y a de commun à tous les items. Cela nous permettrait de voir toutes les interactions qui n'apparaissent pas avec une analyse item par item.

•Question 5: Apports de chacune des méthodes?

Les exemples précédents font ressortir des aspects intéressants de chacune des approches. Face à une analyse de données, peut-être est-il nécessaire de travailler avec toutes ces méthodes pour faire ressortir différents aspects. Au fond, qu'est-ce qu'apporte chacune de ces méthodes ? Ou encore une méthode permet-elle de répondre à une question particulière par rapport aux données ? (autrement dit la manière de poser des questions n'est pas indépendante des outils proposés).

- Réponse de Régis Gras

En réponse à la question posée, qu'est-ce que les différentes analyses peuvent apporter respectivement ? Régis Gras fera la lecture de quelques lignes d'un psychologue, Londeix, psychologue qui a étudié le développement opératoire entre 11-12, 14-15 ans en rapport avec les acquisitions scolaires[1]

1. H. Londeix. "Recherche sur le développement opératoire entre 11-12 et 14-15 ans, ses rapports avec les acquisitions scolaires". Juin 85. IREM de Bordeaux.

et a mis à contribution dans son étude les 3 méthodes (analyse des correspondances, méthodes classificatoire et implicative). Son travail porte sur les performances d'enfants à des tests qui ont été éprouvés, analyse portant sur la classification, la combinatoire, les arrangements, les permutations... la formation de coordonnées spatiales, la logique des propositions... Tout ce qui peut avoir fait partie de la batterie des questionnaires auxquels Piaget soumettait les enfants de la pré-adolescence.

Londeix dit "Le premier facteur des analyses de correspondance répartit les items sur un axe concret-formel. L'opérativité donc devient le principal critère de différentiation des réponses que l'on analyse simultanément ou séparément des différents sous-tests, il apparait toujours en premier". Qu'il le fasse sur les combinaisons ou sur les propositions, c'est toujours une opposition. Le premier facteur est marqué par une opposition qui correspond, par la classification grossière, au stade concret-formel.

"Les analyses hiérarchiques et implicatives révèlent que ce facteur ne traduit pas seulement un accroissement des difficultés, mais dans la plupart des cas le fait que la réussite des items formels est subordonnée à celle des items des stades inférieurs".

C'est ici un autre regard qu'on porte sur les résultats pour voir s'il y a une hiérarchie, une imbrication de ce type, donc un autre type d'analyse. Cette analyse met en évidence que les opérations à un niveau formel mobilisent les opérations qui sont à un niveau inférieur.

"Ce constat laisse penser qu'il existe des filiations logiques entre épreuves et éteint l'hypothèse théorique de l'unique dimensionnalité de développement. Enfin, les analyses classificatoires, méthode Lerman, permettent de retrouver les quasi-classes d'équivalence des items des mêmes stades."

A un stade donné, en fonction des contenus des domaines sur lesquels vont opérer les opérations dont disposent les élèves, il va y avoir des différenciations, c'est ce que Londeix regarde. Et la méthode classificatoire a permis de mettre en évidence des classes d'items.

Jean-Marie Van der Maren conclut cette table ronde en indiquant que si on tente de trouver des choses et non de prouver des choses, il faut utiliser plusieurs méthodes, ce qui permet de mettre de l'ordre, d'organiser---sous une apparence aléatoire. On a alors différents outils à notre disposition : analyse des correspondances, classificatoire--- etc.

Trop souvent, on a peut-être tendance à vouloir prouver alors que nos bases théoriques sont insuffisantes.

ANNEXE 1

PROGRAMME DU COLLOQUE
"Recherches sur l'apprentissage : Analyse des correspondances et méthodes statistiques apparentées"

Lundi le 26 mai 1986

12h00 (local 0-5405)	Accueil des participants et buffet.
13h30 (local 0-4105)	Ouverture par Claude Corbo, Recteur de l'Université du Québec à Montréal.
13h45 (local 0-4105)	Conférence de Brigitte Escofier, Institut de Recherche en informatique et Systèmes Aléatoires de l'Université de Rennes (Institut attaché au CNRS): **La méthode d'analyse des correspondances: grandes lignes théoriques de la méthode, utilisation et dépouillement des résultats.**
16h00	Commentaires, questions, discussion.

Mardi le 27 mai 1986

9h00	Café.
9h30 (local 0-4105)	Conférence de Régis Gras, Institut de Recherche Mathématique de Rennes (Institut attaché au CNRS): **Deux méthodes d'analyse de données didactiques: classification implicative et classification hiérarchique.**
11h30	Commentaires, questions, discussion.
12h00	Repas.
13h30 (local 0-4105)	Communication de Richard Bertrand, INRS-Education: **Méthodes de régression en analyse exploratoire.**
14h30	Café.
14h45 à 16h45	Première série d'ateliers autour d'exemples précis de données brutes traitées par analyse des correspondances provenant de chercheurs-es du Centre.

(local 0-3235) — Atelier 1: Personne-ressource: Brigitte Escofier. Discussion autour de données. **Avantages et limites dans le traitement de ces données par la méthode d'analyse des correspondances.**

(local 0-4105) — Atelier 2: Personne-ressource: Régis Gras. Discussion autour de données. **Avantages et limites de l'utilisation des méthodes classification implicative ou classification hiérarchique.**

Mercredi le 28 mai 1986

9h00 à 10h00
(local 0-4105) — Communication de Bernard Clément, Département de mathématiques, Ecole Polytechnique, Université de Montréal: **Un programme SAS pour faire de l'analyse des correspondances.**

10h15 à 12h00 — Deuxième série d'ateliers (reprise).
Atelier 1.

Atelier 2.

12h00 — Repas.

13h30 — Table-ronde autour de questions qui se posent par rapport aux recherches actuelles sur l'apprentissage et les méthodologies de recherche appropriées. **Intervenants: Régis Gras, Université de Rennes, Brigitte Escofier, Université de Rennes et Jean-Marie Van der Maren, Sciences de l'éducation, Université de Montréal.**

17h00 (local 0-4101) — Cocktail de clôture.

ANNEXE 2

LISTE DES DOCUMENTS DISPONIBLES AU CIRADE ET INDICATIONS SUR D'AUTRES DOCUMENTS INTERESSANTS A SE PROCURER

Fait par Nadine Bednarz, directrice du CIRADE, Juin 1984

En rapport avec l'analyse des correspondances

Livre :

- Jean-Pierre FENELON. **Qu'est-ce que l'analyse des données?** Editeur: LeFonen, 26, rue des Cordelières, 75013. Paris.
  (Tél : (1) 331-17.59)

  Livre s'adressant aux non-mathématiciens, présentant l'analyse des correspondances.

Publications spécifiquement reliées aux méthodes classificatoire et implicative

Articles :

- I.C. LERMAN. "Essai pour une présentation générale de la classification des données" Inf. Sciences humaines, 15e année, no 66, p. 7-40.

  Présentation de la méthode implicative; présentation mathématique succinte des principes sous-jacents.

- I.C. LERMAN, R. GRAS et H. ROSTAM. "Elaboration et évaluation d'un indice d'implication pour des données binaires" (I). Math. Sciences humaines, 19e année, no 74, 1981, p. 5-35.

  Aticle consacré à l'élaboration d'un indice d'implication : modèle mathématique sous-jacent.

- I. C. LERMAN, R.GRAS et H. ROSTAM. "Elaboration et évaluation d'un indice d'implication pour des données binaires" (II) Math. Sciences humaines 19e année, no 75, 1981, p. 5-47.

  Suite de l'article I : présentation du dessin de deux graphes d'implication dans le contexte d'un cas réel fourni par un test mathématique élaboré autour du concept de la symétrie centrale.

Publications internes de l'IRISA
(Institut de recherche en informatique et systèmes aléatoires - Rennes).

- I. C. LERMAN. "Programme de classification hiérarchique"
  (I) Méthode de la vraisemblance des liens
  (II) Méthode de la variance expliquée;
  publication interne de l'IRISA. Juin 1981. 116 pages.

  Présentation du programme de classification hiérarchique de Lerman (4 000 instructions) pour différentes structures du tableau des données; 1ère méthode basée sur la vraisemblance des liens, 2e méthode basée sur le critère de la variation de l'inertie expliquée.

- I.C. LERMAN & P. PETER. "Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de consensus en classification". Rapport de l'IRISA, no 434, juillet 85, 72 pages.

  Présentation d'une méthode très générale de normalisation reliée à la définition d'un indice de similarité entre objets. L'aspect logiciel du calcul, dans différents cas, est présenté avec une description des sous-programmes et de leurs paramètres.

Livre (à se procurer)

- I. C. LERMAN. **Classification et analyse ordinale des données** Dunod, 1981. 760 pages.

  Livre reprenant les méthodes de classification et d'analyse ordinale de données. Le fondement de ces méthodes y est établi d'un point de vue formel, combinatoire et statistique.

En rapport avec l'utilisation des différentes méthodes dans différents domaines

Thèse d'état de Régis Gras

- Régis GRAS. "Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques". Thèse d'état soutenue le 3 octobre 1979, 321 pages.

  Mise au point d'une taxonomie d'objectifs cognitifs -----comprend une analyse factorielle et classificatoire d'un test national d'entrée en 4e en Maths.

Publications reliées aux méthodes classificatoires et implicatives :

Autre thèse :

- Habibullah ROSTAM. " Construction automatique et évaluation d'un graphe "d'implication" issu de données binaires dans le cadre de la didactique des mathématiques", publication interne de l'IRISA, juin 1981, 112 pages.

  Thèse reposant sur la taxonomie d'objectifs cognitifs proposée par R. Gras, adaptation et usage d'un indice statistique d'implication, automatisation du dessin d'un graphe orienté.

- Deux Articles extraits de deux brochures de l'A.P.M.E.P. (Association française des professeurs de Mathématiques de l'enseignement public).

  (1) Régis GRAS, " Analyse d'un questionnaire d'attitudes à l'égard des Mathématiques à l'entrée de 4e expérimentale", p. 161 à 183.
  (Article extrait de la brochure "Analyse de données", Tome I)

  Exemple de traitement d'un questionnaire d'attitudes s'appuyant sur l'analyse en classification hiérarchique de Lerman, et l'analyse factorielle des correspondances.

(2) Régis GRAS. "Analyse factorielle des correspondances entre deux ensembles " (Article extrait de la brochure "Analyse de données", Tome 2), p. 113 à 156.

Présentation de la méthode avec étude d'un exemple.

Pour les brochures complètes, s'adresser à L'A.P.M.E.P.
13, rue du Jura
75013. Paris, FRANCE

Publication interne :

- I. C. LERMAN et P. PETER. "Organisation et consultation d'une banque de "petites annonces" à partir d'une méthode de classification hiérarchique en parallèle." Publication de l'IRISA, 41 pages.

Illustration de la méthode de Lerman sur de "gros ensembles" (nombre important de données).

Articles : exemples d'utilisation des 3 méthodes et des apports de chacune dans le cadre d'une recherche en psychologie

- Hervé LONDEIX. "Les relations entre psychologie génétique et psychologie différentielle" I. Aperçus sur les objectifs et les méthodes". **L'orientation scolaire et professionnelle 1985**, 14, no 1, p. 23-43.

- Hervé LONDEIX. "Les relations entre psychologie génétique et psychologie différentielle" II. L'analyse d'un test opératoire. L'E.C.D.L." **L'orientation scolaire et professionnelle 1985**, 14, no 2, p. 105-144.

Ces deux articles résument une thèse de 3e cycle intitulée "Approches génétiques et différentielles du développement intellectuel, contribution méthodologique à l'analyse des opérations logico-mathématiques", octobre 83. Bordeaux.

• Publication interne de l'IREM de Bordeaux.
Hervé LONDEIX. "Recherche sur le développement opératoire entre 11-12 et 14-15 ans, ses rapports avec les acquisitions scolaires". Juin 1985, 40 pages.

Etude d'un groupe de recherche de l'IREM sur des relations entre le développement de l'activité intellectuelle et l'acquisition des mathématiques au collège.

Autres articles :

• Régis GRAS. " Statistique contemporaine à l'ère informaticienne". Article paru dans le bulletin régional de l'A.P.M.E.P., p. 10 à 16.

Article de réflexion sur le nouveau statut de la statistique à l'ère de l'informatique, exposé fait au séminaire national de didactique des Mathématiques à Paris, 1982.

• Danièle BOISNARD. "Banque d'exercices de Mathématiques" à paraître dans "Petit x."

Travail utilisant la taxonomie de Régis Gras.

Information en rapport avec les logiciels de traitement de données :

• SPAD (système portable pour l'Analyse des données).

SPAD est un outil pour l'analyse exploratoire des grands tableaux de variables nominales ou numériques.
Système d'abonnement (version remise à jour continuellement); version en cours sur micro-ordinateur.

Adresse : CESIA (Centre de statistique et d'informatique appliquées)
Siège social : 82, rue de Sèvres
75007 - Paris
FRANCE

- ADDAD (association pour le développement et la diffusion de l'analyse des données).

Système d'abonnement; version disponible sur micro-ordinateur. Stages organisés : prévision, tableaux multiples, enquête---.
Adresse (stages et courrier) : 22, rue Charcot
75013 - Paris
FRANCE

- Un logiciel de traitement des données par Analyse des correspondances a été conçu par Bernard Clément (Département de Maths Appliquées, Ecole polytechnique) fonctionnant sur S.A.S. (Statistical Analysis System).

Ouvrage de référence : (en rapport avec S. A.S.)

Introduction à S.A.S. par Bernard Clément, 1984 (Ecole Polytechnique).

Livres intéressants à consulter en relation avec les questions et la problématique soulevées au colloque :

- Mark S. ALDENDERFER, Roger, K. BLASHFIELD. "Cluster Analysis" Séries - Quantitative Applications in the Social Sciences. Editors : John L. Sullivan, Richard G. Niemi, Sage Publications, Beverly Hills, London.

- Repères. Essais en éducation. Numéro 6. Numéro thématique Jean-Marie VAN DER MAREN **"Revoir la recherche en éducation : cesser de prédire pour mieux comprendre"**, p. 100 à 140.

Des exemplaires du présent document sont en vente à l'adresse suivante:

Centre Interdisciplinaire de Recherches sur
l'Apprentissage et le Développement en Education
Université du Québec à Montréal
C.P. 8888, Succ. A, G-3125
Montréal, Québec
H3C 3P8

Pour commander, remplir le formulaire de réquisition apparaissant à la dernière page du document.

Publié par le Centre Interdisciplinaire de Recherches
sur l'Apprentissage et le Développement en Education
Université du Québec à Montréal
Conception graphique: Pierrette Prévost